AD_____

Award Number: DAMD17-98-1-8119

TITLE: Genetic Damage Caused by ALU Repeats in Breast Cancer

PRINCIPAL INVESTIGATOR: Prescott Deininger, Ph.D.

CONTRACTING ORGANIZATION: Tulane University Medical Center
New Orleans, Louisiana 70112-2699

REPORT DATE: August 2000

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

20010327 074

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	August 2000	Annual (1 Aug 99 - 31 Jul 00)

4. TITLE AND SUBTITLE
Genetic Damage Caused by ALU Repeats in Breast Cancer

5. FUNDING NUMBERS
DAMD17-98-1-8119

6. AUTHOR(S)
Prescott Deininger, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Tulane University Medical Center
New Orleans, Louisiana 70112-2699

E-MAIL:
pdeinin@tcs.tulane.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***

Retrotransposition by Alu and L1 elements represents 0.2% of all human genetic disease and is likely to be stimulated in tumors. We hypothesize that retrotransposition may be stimulated sufficiently in breast tumors to allow these insertions to make a significant contribution to mutations that may lead to breast cancer progression. Our evolutionary analysis led us to define specific diagnostic sequence differences that differentiate the most recent 2000 Alu inserts from the 1,000,000 older elements. We have developed an anchored PCR strategy that allows us to 'display' small subsets of these recent Alu elements on a gel. We will use this approach to look for de novo Alu insertions in breast tumors. We have recently refined the evolutionary analysis to allow us to detect over one third of all Alu insertions in three subsets that consist of a total of about 150 elements. Each of these subsets can be readily assayed individually. We have also used a retrotransposition reporter system to demonstrate that p53 mutations greatly enhance retrotransposition rates. Because p53 mutations are among the most common breast cancer mutations, this strongly supports our hypothesis.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Breast Cancer	84
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

(4) Table of Contents

(5) Introduction:

This project was based on the hypothesis that early cellular transformation events involved in breast cancer formation might influence the amplification of human Alu repeats. Any increases in Alu amplification, might contribute to further destabilization of the human genome and inactivation of tumor suppressors that could contribute to the progression of breast cancer. At least in sporadic cases, Alu insertions have been shown to contribute to a number of cancers, including at least one case of breast cancer due to inactivation of BRCA2 (1). We have previously shown that only a specific set of subfamilies of Alu elements are actively amplifying in the human genome(2,3). This project combines this information with an anchored PCR procedure we have developed to form displays of the most recently amplified Alu elements. We have demonstrated that this Allele-Specific Alu PCR (ASAP) will effectively display the members of the smallest of the recent Alu subfamilies as bands on an acrylamide gel (5). Our goal is to generalize these procedures to the larger subfamilies and explore various procedures to deal with the larger number of bands expected. We will then use these procedures to compare breast cancer and normal DNA from a number of individuals to determine whether there are new, tumor-specific Alu inserts. This will allow us to determine whether this form of genetic instability plays a role in human breast cancer.

(6) BODY

Goals of Year Two:

- Refinement of Subtraction Technology. Technical development will continue with refinement of the subtraction procedures and tests of the sensitivity of detection of bands and the ability to pool samples in the PCR reactions.
- Preliminary work on tumor samples. Work will begin with existing technology to carry out analysis on tumor samples. We expect to have carried out analysis of the first 10-20 samples in this year. We will use this experience to determine the best approach to generate data in a production mode. This will provide an initial feel for the level of diversity in the displays and a basic characterization of any diversity to determine whether it is caused by insertions. Any evidence of other forms of genomic instability influencing the assay will be assessed at this point and procedures optimized to compensate.

Accomplishments of Year Two:

We have had a number of difficulties with the more technically demanding developments of year 2. First, with numerous attempts to work out the subtraction technology, we have been unsuccessful to date. These experiments have either failed to elicit bands, or in controls in which we started with faint bands, resulted in minimal measurable enrichment and loss of gel resolution (smeary bands). Although we still have some approaches that may improve the subtraction procedure, we have moved most of our effort to the studies below.

By using the data mining procedures from the human genome sequence described in both the Genetica and Genome Research manuscripts, we have uncovered a great deal more detail about the subfamilies of Alu elements that have been most actively inserting in the human genome. In particular, we have identified a subfamily termed Ya5a2 that is an even more recent

and more active version of the Ya5 subfamily (4). More recently, in a study not covered in those manuscripts, we have identified two new subfamilies, Yc1 and Yc2, that represent essentially all of the recent inserts that do not fall into the Ya5 and Yb8 subfamily lineages. The Yc1 and Yc2 subfamilies are very recent in origin and about 70% of them are polymorphic in the human lineage (compared to 30% of Ya5). In addition 3 out of 15 human diseases caused by Alu insertions are from the Yc lineage. There are only about 50 Yc subfamily members in the human lineage, making this subfamily, like the Ya8 and Ya5a2 lineages manageable without subtraction procedures using our PCR technique. Furthermore, these small subfamilies represent about 6 out of 16 disease-causing Alu insertions, making it likely that we can cover more than a third of active Alu insertions just by studying these minor subfamilies. Therefore, we are turning most of our attention to amplifying each of these subfamilies effectively.

The second problem we have had this year is that we have found our breast cancer samples less than ideal to work with. Some have had only very small amounts of tissues and a few have probably just been stored too long, but we have had a general problem getting good clean DNA from the samples that cuts with restrictions enzymes and behaves well in our anchored PCR technology. We have obtained ten new breast cancer DNA samples from Dr. Steven Hill, and are beginning to work with those samples at this point.

We have also just generated a very important observation that is relevant to the goals of our project. We have been using a reporter gene system that allows us to measure the rate of L1 retrotransposition from a specific L1 element that we introduce into cells. When this L1 undergoes retrotransposition, it activates a neomycin resistance gene. We can then use G418 to kill cells without the neomycin resistance and count resistant colonies as a measure of retrotransposition rate. We carried out an experiment where we transiently transfected the L1 reporter system with control plasmid, or various p53 mutants. Both p53 mutants increased the number of colonies detected significantly and one increased the rate of retrotransposition by well over an order of magnitude. This is direct evidence that p53 mutations are capable of increasing the retrotransposition rate. Because we believe that Alu works by pirating the retrotransposition apparatus of L1, it seems likely that both L1 and Alu amplification increases greatly in the presence of p53 mutations. Because p53 mutations are one of the most common mutations in breast cancer, this suggests that this process may be of increased importance in breast cancer. It is our intention to repeat these studies during the next year using MCF7 as a human breast cancer cell line and to introduce the major p53 mutations that are found associated with breast cancer in order to get a more meaningful estimate of the increased amplification expected in actual breast tumors.

(7) Key Research Accomplishments

Year 1

- Establishment of optimum conditions for amplification of the most recent subfamilies of Alu inserts
- Obtaining clear displays of the Ya8 subfamily on acrylamide and agarose gels which allow the isolation of insertion polymorphisms between different individuals.
- Demonstrating the use of modified primers that display subsets of the Ya5 elements that will allow at least a substantial portion of Ya5 inserts to be studied.

Year 2

- Identification of the youngest, most active Alu subfamilies that can be amplified and displayed directly without the use of subtraction protocols.
- Demonstrating the influence of p53 on the rate of retrotransposition.

(8) Reportable Outcomes

At this point, we have published one review article that acknowledges this grant support, and have two research publications in press (appended). These studies outline the importance of Alu insertion to genetic instability, demonstrate the basic PCR technology developed in this project and present some of the new subfamily analysis that will allow us to focus on the most active subset of elements.

(9) Conclusions

This project requires the further development of existing techniques to answer the question of whether human mobile elements contribute significantly to genomic instability in breast cancer. We are in a position to make a major contribution to this area in the next year, although it will be difficult to complete the major study intended.

Our finding, using a reporter gene system to measure L1 retrotransposition, shows that p53 mutations greatly increase retrotransposition rates. Although not a direct measure of this activity in breast tumors, this is a very important observation that implicates mobile elements in instability in breast tumors.

(10) Reference List

1. P.L. Deininger, M.A. Batzer, *Mol Genet Metab* **67, 183** (1999).

2. M. Batzer, *et al, Nucleic Acids Res.* **19**, 3619 (1991).

3. P. Deininger, V. Slagel, *Mol.Cell.Biol.* **8**, 4566 (1988).

4. Roy et al, *Genome Res.* (in press)

5. Roy et al, *Genetica* 107:149-161 (1999)

APPENDIX

one reprint for:

Roy et al, *Genome Res.* (in press)

Roy et al, *Genetica* 107:149-161 (1999)

RECENTLY INTEGRATED HUMAN ALU REPEATS: FINDING NEEDLES IN THE HAYSTACK

Keywords: Alu insertion polymorphisms, anchored-PCR, comparative genomics, computational biology

Astrid M. Roy[1+], Marion L. Carroll[2+], David H. Kass[3+], Son V. Nguyen[2+], Abdel-Halim Salem[2,^], Mark A. Batzer[2], Prescott L. Deininger[1,4]*

[1]Tulane Cancer Center, SL-66, Department of Environmental Health Sciences, Tulane University - Medical Center, 1430 Tulane Ave., SL-66, New Orleans, Louisiana 70112.

[2]Departments of Pathology, Biometry and Genetics, Biochemistry and Molecular Biology, Stanley S. Scott Cancer Center, Neuroscience Center of Excellence, Louisiana State University Health Sciences Center, 1901 Perdido Street, New Orleans, Louisiana 70112.

[3]Department of Biology, 316 Mark Jefferson, Eastern Michigan University, Ypsilanti, MI 48197.

[4]Laboratory of Molecular Genetics, Alton Ochsner Medical Foundation, 1516 Jefferson Highway, New Orleans, Louisiana 70121.

[^]Present address: Department of Anatomy, Faculty of Medicine, Suez Canal University, Ismailia, Egypt.

[+]These authors contributed equally to this manuscript.

* Corresponding author.
Mailing address: Tulane Cancer Center
Tulane University - Medical Center
1430 Tulane Ave., SL-66
New Orleans, LA 70112
Phone: (504) 988- 6385
Fax: (504) 588-5516
Internet: PDEININ@TCS.TULANE.EDU

ABSTRACT

Alu elements undergo amplification through retroposition and integration into new locations throughout primate genomes. Over 500,000 Alu elements reside in the human genome, making the identification of newly inserted Alu repeats the genomic equivalent of finding needles in the haystack. Here, we present two complementary methods for rapid detection of newly integrated Alu elements. In the first approach we employ computational biology to mine the human genomic DNA sequence databases in order to identify recently integrated Alu elements. The second method is based on an anchor-PCR technique we term Allele-Specific Alu PCR (ASAP). In this approach, Alu elements are selectively amplified from anchored DNA generating a display, or "fingerprint", of recently integrated Alu elements. Alu insertion polymorphisms are then detected by comparison of the DNA fingerprints generated from different samples. Here, we explore the utility of these methods by applying them to the identification of members of the smallest previously identified subfamily of Alu repeats in the human genome termed Ya8. This subfamily of Alu repeats is composed of about seventy elements within the human genome. Approximately fifty percent of the Ya8 Alu family members have inserted in the human genome so recently that they are polymorphic making them useful markers for the study of human evolution.

INTRODUCTION

Alu repeats are the most successful class of mobile elements in the human genome. Alu elements spread through the genome via an RNA mediated amplification mechanism termed retroposition reviewed in Deininger & Batzer, 1993. There are over 500,000 Alu elements in the human genome, which have clearly played a major role in sculpting and damaging the genome. Alu elements have contributed to genetic disease, both by the disruption of genes through the insertion of newly retroposed elements and by recombination between Alu elements (reviewed in
(Deininger & Batzer 1999). Previous estimates indicate that retroposition of Alu elements contributes to approximately 0.1% of human genetic diseases and recombination between Alu repeats contributes to another 0.3% of genetic diseases (Deininger & Batzer 1999). Therefore, the spread of the Alu family of mobile elements has generated a significant amount of human genomic variation as well as diseases through recombination-based fluidity as well as insertional mutagenesis.

Alu repeats are distributed rather haphazardly throughout the human genome. Alu elements began expanding in the ancestral primate genomes about 65 mya (Shen et al. 1991) reaching a peak amplification between 35 and 60 mya. Presently, Alu elements amplify at a rate that is 100 fold lower than the maximum rate, with an estimate of one new Alu insert in every 100-200 births (Deininger & Batzer 1993, 1995). Evolutionary studies have demonstrated that the majority of evolutionarily recent Alu inserts have specific diagnostic sequence mutations (Deininger & Batzer 1993; Deininger & Batzer 1995). These mutations have accumulated in Alu elements throughout primate evolution resulting in a hierarchical subfamily structure, or lineage, of Alu repeats. The mutations

facilitate the classification of Alu elements into different subfamilies, or clades, of related elements that share common diagnostic mutations reviewed in (Batzer et al. 1993; Batzer & Deininger 1991; Batzer et al. 1996a). Almost all of the recently integrated Alu repeats within the human genome belong to one of four closely related subfamilies: Y, Ya5, Ya8, and Yb8, with the majority being Ya5 and Yb8 subfamily members. Collectively, these subfamilies of Alu elements comprise less than 10% of the Alu elements present within the human genome with the Ya5/8 and Yb8 subfamilies collectively accounting for less than half of a percent of all Alu elements. These evolutionary recent Alu insertions are useful for human population studies, since there appears to be no specific mechanism to remove a newly inserted Alu repeat and the elements are identical by descent with a known ancestral state (Batzer et al. 1991; Batzer et al. 1994a; Batzer et al. 1996b; Stoneking et al. 1997; Perna et al. 1992).

Previously, it has been technically impossible to determine the full impact of mobile elements on the human genome. The identification of newly inserted Alu elements has been very difficult due to the complexity of detecting one new Alu insertion in a cell that already has 500,000 pre-existing Alu elements. We have previously utilized laborious library screening and sequencing strategies to isolate relatively small numbers of Alu insertion polymorphisms (Arcot et al. 1995a) (Arcot et al. 1995b; Arcot et al. 1995c; Batzer & Deininger 1991; Batzer et al. 1990; Batzer et al. 1991; Batzer et al. 1995), as well as investigating rare 300 bp restriction fragment length polymorphisms ((Kass et al. 1994)). This makes these studies the genomic equivalent of the search for needles in the haystack. In this paper we discuss two alternative methods that overcome the inherent difficulties in these experiments, making these studies manageable. First, the

availability of large quantities of human genomic DNA sequence provided by the Human Genome Project facilitates genomic database mining for recently integrated Alu elements. This approach should prove useful in determining the chromosome-specific and genome wide dispersal patterns of mobile elements, as well as for the identification of polymorphic mobile element fossils to apply to the study of human population genetics and primate comparative genomics. Secondly, we have developed a PCR-based method that we term Allele-Specific Alu PCR (ASAP). This technique allows us to take advantage of the subfamily-specific diagnostic mutations within Alu mobile elements to isolate and display recently integrated Alu repeats from different DNA samples; allowing for direct comparisons of the Alu content of different genomes or different cells from an individual.

MATERIALS AND METHODS

Cell lines and DNA samples

The cell lines used to isolate human DNA samples were as follows: human (*Homo sapiens*), HeLa (ATCC CCL2); chimpanzee (*Pan troglodytes*), Wes (ATCC CRL1609); gorilla (*Gorilla gorilla*), Ggo-1 (primary gorilla fibroblasts) provided by Dr. Stephen J. O'Brien, National Cancer Institute, Frederick, MD, USA. Cell lines were maintained as directed by the source and DNA isolations were performed using Wizard genomic DNA purification (Promega). Human DNA samples from the European, African American and Greenland Native population groups were isolated from peripheral blood lymphocytes (Ausubel et al. 1996) that were available from previous studies (Stoneking et al. 1997). Egyptian samples were collected from throughout the Nile

River Valley region and DNA from peripheral lymphocytes was prepared using Wizard genomic DNA purification kits (Promega). Human DNA used for ASAP was isolated from peripheral lymphocytes utilizing the Super-Quick Gene method (Analytical Genetic Testing Center).

Computational analyses

A schematic overview summarizing the computational analyses of recently integrated Alu elements is shown in Figure 1. Initial screening of the GenBank non-redundant and high throughput genomic sequence (HTGS) databases was performed using the Basic Local Alignment Search Tool (BLAST) (Altschul et al. 1990) available from the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/). The database was searched for exact complements to the oligonucleotide 5'-ACTAAAACTACAAAAAATAG-3' that is an exact match to a portion of the Alu Ya8 subfamily consensus sequence containing unique diagnostic mutations. Sequences that were exact complements to the oligonucleotide were then subjected to more detailed annotation. A region composed of 1000 bases of flanking DNA sequence directly adjacent to the sequences identified from the databases that matched the initial GenBank BLAST query were subjected to annotation using the RepeatMasker2 program from the University of Washington Genome Center server (http://ftp.genome.washington.edu/cgi-bin/RepeatMasker). This program annotates the repeat sequence content of individual sequences from humans and rodents.

Primer design and PCR amplification

PCR primers were designed from flanking unique DNA sequences adjacent to individual Ya8 Alu elements using the Primer3 software (Whitehead Institute for Biomedical Research, Cambridge, MA, USA) (http://www.genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi). The resultant PCR primers were screened against the GenBank non-redundant database for the presence of repetitive elements using the BLAST program, and primers that resided within known repetitive elements were discarded and new primers were designed. PCR amplification was carried out in 25 μl reactions using 50-100 ng of target DNA, 40 pM of each oligonucleotide primer, 200 μM dNTPs in 50 mM KCl, 1.5 mM $MgCl_2$, 10 mM Tris-HCl pH 8.4 and Taq® DNA polymerase (1.25 u) as recommended by the supplier (Life Technologies). Each sample was subjected to the following amplification cycle: an initial denaturation of 2:30 min at 94°C, 1 min of denaturation at 94°C, 1 min at the annealing temperature, 1 min of extension at 72°C, repeated for 32 cycles, followed by a final extension at 72°C for 10 min. Twenty microliters of each sample was fractionated on a 2% agarose gel with 0.25 μg/ml ethidium bromide. PCR products were directly visualized using UV fluorescence. The sequences of the oligonucleotide primers, annealing temperatures, PCR product sizes and chromosomal locations are shown in Table 1. Phylogenetic analysis of all the Alu elements listed in Table 1 was determined by PCR amplification of human and non-human primate DNA samples. The human genomic diversity associated with each element was determined by the amplification of 20 individuals (160 total chromosomes) from each of four populations (African-American, Greenland Native, European and Egyptian). The chromosomal location of Alu repeats identified from clones that had not

been previously mapped was determined by PCR amplification of National Institute of General Medical Sciences (NIGMS) human/rodent somatic cell hybrid mapping panel 2 (Coriell Institute for Medical Research, Camden, NJ).

Allele-Specific Alu PCR (ASAP)

We used a modification of the IRE-Bubble PCR method (Munroe et al. 1994), utilizing the same amplification (anchor) primer, but altering the annealed anchor/linker primers. The annealed linkers formed a Y instead of a bubble to avoid end-to-end ligation. Also, instead of blunt-end digestion, genomic DNA was digested with *Mse*I, that cuts 5'-T'TAA-3' and does not cut in the Alu consensus. Otherwise the genomic-anchor ligations were prepared according to (Munroe et al. 1994). The annealed linker primers are: MSET: 5'-TAGAAGGAGAGGACGCTGTCTGTCGAAGG-3' and MSEB: 5'-GAGCGAATTCGTCAACATAGCATTTCTGTCCTCTCCTTC-3'. The amplification (linker) primer is: LNP: 5'GAATTCGTCAACATAGCATTTCT-3'. We placed an *Eco*RI site at the 5' end of the primer for the option of cloning PCR products into cloning sites of common vectors. No bands are observed on a gel when this primer is used alone with the anchored template at an annealing temperature of 55° C.

Unless otherwise noted, PCR conditions (for all ASAP reactions) were performed in 20 µl using a Perkin-Elmer 9600 thermal cycler with the following conditions: 1x Promega buffer, 1.5 mM $MgCl_2$, 200 µM dNTPs, 0.25 µM primers, 1.5 U Taq polymerase (Promega) at 94°C-2 m; 94°C-20 s, 62°C-20 s, 72°C-1 m 10 s, for 5 cycles; 94° C-20 s, 55° C-20 s, 72° C-1 m 10 s, for 25 cycles; 72° C - 3 m. Nested Alu primers were used that move along the Alu in an upstream direction as follows: ASII (Ya5-

specific): 5'-CTGGAGTGCAGTGGCGG-3'; HS18R (Ya8-specific): 5'-CTCAGCCTCCCAAGTAGCTA-3'; HS16R (Ya8 (HS-specific): 5'-CGCCCGGCTATTTTTGTAG-3'.

The ASII primer has Ya5 diagnostic nucleotides (present in both Ya5 and Ya8 subfamilies). In the first round of PCR, stock genomic DNA (2.4 ng anchored DNA) was used as the template. For subsequent rounds of amplification, PCR products were purified through microcon-30 (Amicon) columns using two centrifuge spins following the addition of 400 μl of water. For the second round of amplification, 1 μl of microcon-purified first round PCR reaction was used as the template, and for the third round 1 μl of microcon-purified second round PCR products was used. For display analysis (see below) the PCR products were "equalized" in volume following microcon purification.

Display of Anchor-Alu PCR products

Third round PCR was performed utilizing a 5' end-labeled primer incorporating [γ-^{32}P] ATP (Amersham) with T4 polynucleotide kinase (New England BioLabs). PCR conditions were as above with the exception of using 0.188 μM of each Ya8 and LNP cold primers and 0.075 μM of end-labeled Ya8 primer. Anchor-PCR and end-labeled molecular weight markers ($\phi\chi$174 DNA digested with *Hinf*I; Promega) were separated by electrophoresis on denaturing 5% Long Ranger (AT Biochem) gels and examined by autoradiography following exposure to Amersham Hyperfilm at room temperature. Individuals from different ethnic groups were utilized to select for variants that may have been the result of a recent Alu insertion event (polymorphism).

Verification of Bands as Ya8 products

Gels were aligned to autoradiographs by either small cuts in various parts of the gel, or placement of low-level radioactive dye on the gel prior to re-exposure. Bands were then sliced out of the gels, placed in 200 μl of water and eluted by heating at 65°C for 15 minutes. Samples were re-amplified with third round PCR primers, cloned and sequenced as described above. Following verification that these bands were amplified by the third round primer pair, new nested oligonucleotides based on the flanking unique sequences were designed to move, by PCR, downstream through the Alu element to the opposite flank. Annealing temperatures were varied based on the Tm of the oligos. Generally two or three rounds of PCR were utilized to obtain the 3' flanking sequences of the Alu. These PCR products were also cloned and sequenced in the same manner.

RESULTS

We present two complementary approaches that facilitate rapid detection of newly inserted Alu elements from the human genome. First, computational analyses of human genomic DNA sequences from the GenBank database are used in the identification of recently integrated Alu elements. Second, allele-specific PCR amplification is used for the selective enrichment of young Alu elements. To compare and contrast these two approaches, we present the data obtained when these methods are applied to the identification of members of the Ya8 Alu subfamily, the smallest previously reported subfamily of Alu repeats in the human genome.

Copy number and sequence diversity

In order to estimate the copy number of Ya8 Alu family members, we determined the number of exact matches to our subfamily specific oligonucleotide query sequence as a proportion of the human genome that had been sequenced in the non-redundant database. 27 matches to the subfamily specific query sequence were obtained from the non-redundant database. Upon further sequence annotation using the RepeatMasker2 web site, five matched the Ya8 Alus previously sequenced in our laboratories (Batzer et al. 1990; Batzer & Deininger 1991; Batzer et al. 1995). Eight of the elements identified in the search were classified as Alu Sx subfamily members, and two matched to the TPA 25 Ya8 Alu family member. A total of 13 independent Ya8 Alu elements were identified from the search of the non-redundant database that were not sequenced as part of a project to specifically identify recently integrated Alu elements. The non-redundant database contained 45.3 % human DNA sequences for a total of 590,140,703 bases of human sequence on the date of the search. The estimated size of the Ya8 subfamily is (3 x 10^9 bp / 590,140,703 bp) x 13 unique Ya8 matches = 66 Ya8 subfamily members. This estimate compares favorably with that of 50 previously reported based upon library screening, restriction digestion or Southern blotting (Batzer et al. 1995). An additional 6 matches to the Ya8 subfamily query sequence were identified in the high throughput genomic sequence database (HTGS). One of these elements was an Alu Sq subfamily member, while a second element was a duplicate copy of Ya8NBC60. PCR analyses of two elements identified in the high throughput database, Ya8NBC7 and Ya8NBC16 (GenBank accession numbers AL109937 and AC008944), were inconclusive and these elements were eliminated from further analysis. These two elements were identified

from low pass first sequence runs in the HTGS database. It is not surprising that the PCR analyses failed, since the DNA sequences are presumably lower quality than finished DNA sequences contained in the non redundant database. However, two additional Ya8 Alu repeats (Ya8NBC8 and Ya8NBC15) were identified in the HTGS database and subjected to analysis.

A comparison of the nucleotide sequences of all of the Ya8 Alu family members is shown in Figure 2. In order to determine the time of origin for the Ya8 subfamily we divided the nucleotide substitutions within the elements into those that have occurred in CpG dinucleotides and those that have occurred in non-CpG positions. The distinction between types of mutations is made because the CpG dinucleotides mutate at a rate that is about 10 times faster than non-CpG positions (Labuda & Striker 1989; Batzer et al. 1990) as a result of the deamination of 5-methylcytosine (Bird 1980). A total of 14 non-CpG mutations and 8 CpG mutations occurred within the 14 Alu Ya8 subfamily members reported. Using a neutral rate of evolution for primate intervening DNA sequences of 0.15% per million years (Miyamoto et al. 1987) and the non-CpG mutation rate of 0.413% (14/3388 using only non-CpG bases) within the 14 Ya8 Alu elements yields an estimated age of 2.75 million years old for the Ya8 subfamily members. This estimate of age is somewhat higher than the 660,000 years previously reported (Batzer et al. 1995). However, the previous study of Ya8 Alu family members involved only four elements making the calculated age more subject to random statistical fluctuation. This estimate is also consistent with the expansion of a family of mobile elements that began around the time humans and African apes diverged, which is thought to have occurred 4-6 million years ago (Miyamoto et al. 1987).

Inspection of the nucleotide sequences flanking each Ya8 Alu family member shows that all of the elements were flanked by short perfect direct repeats (Figure 3). The direct repeats ranged in size from 3-17 nucleotides. These direct repeats are fairly typical of recently integrated Alu family members. Two of the Alu Ya8 Alu family members contained 5' truncations (Ya8NBC2 and Ya8NBC11). Since Ya8NBC2 and Ya8NBC11 are both flanked by perfect direct repeats the truncations in these elements probably occurred as a result of incomplete reverse transcription or improper integration into the genome rather than by post-integration instability. All of the Ya8 Alu family members had oligo-dA rich tails that ranged in length from a minimum of 12 nucleotides to over 40 bases in length. It is also interesting to note that the 3' oligo-dA rich tails of several of the elements (Ya8NBC2, Ya8NBC3, Ya8NBC4, and Ya8NBC8) have accumulated random mutations beginning the process of the formation of simple sequence repeats of varied sequence complexity. The oligo-dA rich tails and middle A rich regions of Alu elements have previously been shown to serve as nuclei for the genesis of simple sequence repeats (Arcot et al. 1995b).

Phylogenetic distribution, and chromosomal location

The phylogenetic distribution of each Ya8 Alu element was determined by amplifying genomic DNA from two non-human primates (common chimpanzee and gorilla). All of the Ya8 Alu family members except Ya8NBC10 were absent from the genomes of non-human primates. This suggests that the majority of these elements dispersed within the human genome sometime after the human and African ape divergence and that less than 7% (1/14 elements) of the randomly sequenced Ya8 Alu

subfamily members reside in non-human primate genomes. The chromosomal location of each Ya8 Alu element was taken directly from the GenBank database entry or determined by PCR amplification of human/rodent monochromosomal hybrid cell line DNA samples.

Human genomic diversity

In order to determine the human genomic variation associated with each of the Ya8 Alu family members we subjected a panel of human DNA samples to PCR amplification (Table 2). The panel was composed of 20 individuals of European origin, African Americans, Greenland Natives and Egyptians for a total of 80 individuals (160 chromosomes). Using this approach four of the 14 (Ya8NBC8, Ya8NBC10, Ya8NBC14 and Ya8NBC15) Alu Ya8 subfamily members were monomorphic for the presence of the Alu element suggesting that these elements integrated in the genome prior to the radiation of modern humans from Africa. Three of the elements (Ya8NBC2, Ya8NBC13 and Ya8NBC17) appeared heterozygous in all of the individuals that were analyzed suggesting that they had integrated into undefined repetitive elements within the human genome. However, the remaining seven elements were polymorphic for the presence of an Alu repeat within the genomes of the test panel individuals (Table 2). The unbiased heterozygosity values (corrected for small sample sizes) for these polymorphic Alu insertions were variable, and approached the theoretical maximum in several cases. This is quite interesting since the maximum uncorrected heterozygosity for these biallelic elements is 50% and suggests that these Alu insertion polymorphisms will make excellent markers for the study of human population genetics. In addition, 50% of the randomly

identified Ya8 Alu family members are polymorphic. These result suggest that the Ya8 subfamily is younger than either the Ya5 (from which Ya8 was derived) or Yb8 Alu subfamilies, since only 25% of the members of these Alu subfamilies are polymorphic in the human genome (Batzer et al. 1995).

Allele-Specific Alu PCR (ASAP).

Although database screening is extremely efficient for identifying recent Alu elements, it will not allow identification of new elements from genomes not included in the sequencing efforts. Our primary objective with the ASAP technique is to rapidly identify newly inserted Alu elements from a background of 500,000 older Alus. To accomplish this feat, we utilized a modification of the IRE-bubble PCR technique (Munroe et al. 1994). The procedure utilizes an anchored PCR strategy (Figure 4) in which genomic DNA is cleaved with an enzyme that does not cleave within the Alu repeat. The modified anchor is then ligated to the fragment ends. This anchor will only allow PCR amplification if a primer first primes within the fragment and replicates across the linker eliminating any problems with amplification from anchor to anchor. We take advantage of the base changes that identify the younger Alu subfamily members (Batzer et al. 1996a; Batzer & Deininger 1991). In addition, this allows the selective enrichment for a smaller fraction of the Alu elements from the genome, as there are only 1000 Ya5 and 1000 Yb8 Alu repeats and approximately 70 Ya8 Alu family members in the human genome (Batzer et al., 1995). We gain the specificity for the recent inserts by using a PCR primer that matches the particular Alu subfamily with the diagnostic positions at its 3' end. Each amplification will extend from a specific Alu subfamily member through its

upstream flanking sequences to the randomly located flanking restriction site. The numerous older Alu repeats have accumulated many mutations and may compete for the PCR primers with the Ya5/8 elements. Therefore, although the first amplification provides a great deal of subfamily specificity, we then carry out a 'nested' reaction using a second allele-specific primer to improve the specificity, followed by a third round with another allele-specific primer. In theory, we can utilize primers for each of the 5-8 diagnostic mutations in a subfamily.

In the example presented in this paper, we focused our attention on the identification and display of the lower copy number Alu Ya8 subfamily. Also to better display the results we used nested primers in the upstream direction of Ya8 to avoid amplification problems through the A-rich tail. Using the primers described in the Materials and Methods section, by the third round of PCR we were able to visualize discreet DNA fragments on an agarose gel (data not shown). The size range of these fragments appeared to be between 150 bp and 800 bp. To enhance this display, we chose an alternative method of electrophoretic separation and end-labeled the nested primer to further minimize background (see below). To verify these were Ya8 repeats we directly cloned the third round PCR products, and sequenced them. Partial or complete sequences of these products using vector primers in both directions, demonstrated all twelve clones to be amplified by the Alu-anchor primer pair, although in one case the unique linker sequence was imprecise. All these elements contained the Ya5/8 diagnostic nucleotides (there were no further upstream diagnostics to declare these as Ya8 elements).

For eight of the twelve isolated clones, there were between 12 and 18 unique nucleotides between the linker and the Alu (or truncated Alu) sequences. Since Alu

elements preferentially insert into A-T rich regions (Daniels & Deininger 1985) and *Mse*I cuts at the sequence TTAA, then this result is not surprising. The advantage of using *Mse*I for the restriction digestion is that most of the Alu-linker products are small enough to be amplified. Although, it would be difficult to perform nested PCR in the opposite direction with those few A-T rich nucleotides, searching GenBank using the BLAST program with the obtained flanking unique DNA sequences as the query may in some cases identify the rest of the genomic sequence for each Alu element. This will provide the Alu location with both its flanking sequences. Flanking unique sequence primers can then be designed and the Alu polymorphism can then be confirmed using other human DNA sources. Once the polymorphism is confirmed subsequent population studies can be performed.

Display and rapid identification of Ya8 associated variants.

To alleviate the need for testing every Ya8 element obtained by this assay, we chose to end-label the third round nested PCR primer to enable a display of individual Ya8 repeats following electrophoretic separation and autoradiography. Observed variations may be due to primer mismatch, genomic rearrangements, small insertion/deletions or Alu based insertion/deletions (I/D).

We carried out the procedure with four different individuals to discern which bands represent variants (Figure 5), and to effectively display variants as DNA fingerprints. We obtained about 40 bands per individual from a single reaction; among the four individuals analyzed, about one half appeared variant (Figure 5). We have developed a potent method for the generation of Ya8 associated DNA fingerprints that is

in reasonable agreement with the database mining approach and seems to display the majority of Alu subfamily members. This necessitated addressing what proportion of the fragments generated were the result of the presence of a Ya8 Alu element and whether the lack of the same band in another individual represented an Alu insertion polymorphism. We chose twelve bands to re-amplify and verify as Ya5/8 elements. Those bands that appeared variant were analyzed for Alu insertion polymorphisms. Other bands were selected for future testing of dimorphisms as these individual Ya8 elements may display variation among other people/populations. Occasionally upon re-amplification from the isolated band we obtained background products and therefore generally more than one clone was sequenced. Of the twelve isolated bands (Figure 5) nine were verified as precisely amplified HS16R-LNP products. Two others each contained a Ya5/8 Alu, one randomly amplified by HS16R (anc-8) in lieu of the linker primer, while anc-3 contained sequences downstream of HS16R. Anc14 apparently was an amplified J (PS) Alu element (data not shown). Therefore, this demonstrates the majority of the bands visualized on the autoradiograph are AluYa5/8 repeats and most probably Ya8. The numerous bands at about 178 nt coincides with our previous finding that many of the products will have between 12 and 18 unique sequences. Of the nine bands that we attempted to obtain the opposite flank by nested anchored PCR, we reached the opposite (downstream) flank of the Alu for three of them (anc-5, anc-6, anc-4). In some cases the amount of unique sequence was too small to employ nested primers and in some cases there was a high level of A-T richness. In one case we merely got a non-specific product. All three sequences obtained were authentic Ya8 Alu elements based on the diagnostic nucleotide positions and the high level of conservation of the

sequence in relation to the consensus. This demonstrates the successful nature of our protocol to select for this subfamily of repeats amongst a large background of Alu repeats.

When "crossing" the anc-5 Alu by nested PCR using four individuals (not all identical to Figure 5), we found a correspondence between the generation of a distinct band among the individuals that also had the anc-5 band on an autoradiograph. However, we obtained a short 3' flank of twelve nucleotides that proved difficult in amplifying DNA from various individuals with unique flanks. It is still possible that this variant represents an I/D event. Besides anc-5, anc-6 also appeared polymorphic on the autoradiograph, although anc-4 did not. However, since we had both flanks, for these Alu elements, we developed primers to rapidly assess various individuals for an insertion variant. For anc-6, one of a few different primer sets worked well, yielding the band of expected size, although also generating a few non-specific bands. However, a band was present for eleven unrelated individuals analyzed (data not shown), including those observed on the autoradiograph, suggesting the anc6 polymorphism was not the result of an I/D variant. In addition, this band was absent in the chimpanzee possibly indicating the absence of the Alu, or perhaps primer mismatch due to nucleotide divergence. Although anc-4 was not variant on the autoradiograph, we tested 13 individuals of various ethnic backgrounds for an I/D event and observed it to be monomorphic. Although we have not verified any of the displayed variants to be the result of an Alu insertion, this potential remains, as we observed Ya8 elements to be highly polymorphic, and all the bands, but one, analyzed were Ya8 repeats.

DISCUSSION

In this manuscript we present an analysis of the smallest defined subfamily of Alu repeats located within the human genome termed Ya8. This subfamily of Alu elements was derived from the Ya5 subfamily of Alu elements. The Ya5 subfamily is composed of approximately 1000 members and has largely integrated into the human genome sometime after the human African ape divergence. The main reasons that support the more recent origin of the Ya8 subfamily are the accumulation of three additional diagnostic mutations as compared to the Ya5 subfamily and the lower copy number for the Ya8 subfamily. However, it is important to note that a small percentage of the Ya8 Alu family members are also located in the genomes of non-human primates. These data indicate that although the copy number of the Ya8 subfamily is small, that it may have begun to propagate much earlier in human evolution than was previously thought. In fact, the Ya8 subfamily may have amplified from an allelic variant of the Ya5 subfamily of Alu elements that was not as efficient at mobilization as the Ya5 source gene.

The ability to detect a handful of Alu repeats from the background of several hundred thousand Alu elements in the human genome is impressive. The application of computational biology to the analysis of large multigene families such as Alu repeats offers the potential to address a number of new questions in comparative genomics as an increasing proportion of the human genome is sequenced. Studies of the present, as well as historical, integration patterns of mobile elements in the human genome may begin to be addressed. In addition, the patterns of diversity generated by the integration of mobile elements into the human genome may be analyzed at a scale that was previously unimaginable. These types of studies will shed new insight into the relationships

between different types of mobile elements in the human genome, integration site preferences, impact, and the biological properties of these elements.

The development of the ASAP technique facilitated the display a subset of Ya8 Alu elements from a large and complex background. The preferential isolation of the young Alu elements, as demonstrated here, enhances the identification of recent Alu insertion events in the genome. We focused our efforts on the smallest known defined subfamily of Alu repeats to best address issues of sensitivity of the display of individual elements. One of the advantages of this technique is its flexibility. The alteration of the restriction enzyme used for digestion of genomic DNA selects distinct subsets of Alu elements within a particular subfamily, since this technique preferentially amplifies products that range from 200 and 800 bp in size. In addition, modifications to the ASAP technique, such as the use of a less frequent restriction endonuclease, may allow for a display of subsets of the larger groups of Alu repeats such as Ya5 elements. Alternatively, the use of primers that select for subfamily "subgroups" may also be used to reduce the complexity of the resultant display by decreasing the number of PCR products. Although we focused on Ya8 Alu elements due to their low copy number, the young Yb8 Alu subfamily is another alternative for ASAP with an estimated copy number of only 1000 elements (Batzer et al. 1995; Zietkiewicz et al. 1994) and some polymorphic members (Hutchinson et al. 1993; Hammer 1994; Arcot et al. 1998). We have previously demonstrated the isolation of young Alu elements (based on sequence identity to a consensus) using a Yb8 diagnostic primer, and a generic Alu as an anchor in the amplification reaction, that can be profiled with minimal background (Kass et al. 1996). It is conceivable that variations on the anchored-Alu PCR technique can be

employed to rapidly localize individual elements from all three subfamilies of young Alu elements.

Once the flanking sequences of the young Alu elements are obtained, the PCR strategy or computational methods reported here can be employed to trace polymorphisms that have resulted from recent Alu insertions and are not yet fixed in human populations. The anchored-Alu PCR approach facilitates rapid identification of young elements by displaying the amplification products, but will also increase the potential for selecting only those mobile element fossils that exhibit presence/absence variation. Selection in this manner also shifts the spectrum for new elements toward the elements that are lower frequency and less likely to be held in common between individuals or populations. Therefore, this approach should prove to be quite useful for the ascertainment of mobile element fossils to address questions about more recent human diversifications. In contrast, the identification of mobile element fossils using computational biology affords the opportunity to identify multiple frequency classes of Alu elements that are shared at different geographic levels within the human population.

The ASAP method's strength comes from its ability to isolate a subset of interspersed repeat sequences from different DNA sources and compare them at the same time. In other words, this approach is not limited to Alu elements, but may be used with other SINEs (from other organisms) or even long interspersed elements (LINEs) or for that matter any repeated DNA sequence family that has a defined subfamily structure. A second potential application would be the use of ASAP to monitor genomic instability associated with different forms of cancer by providing a multilocus monitoring system.

Due to its high flexibility the ASAP technique has an enormous range of potential applications.

Mobile element fossils have proven to be simple powerful tools for tracing the origin of human populations (Perna et al. 1992; Batzer et al. 1994b; Batzer et al. 1996a; Stoneking et al. 1997). These elements should also prove quite useful to the forensic community as paternity identity testing reagents (Batzer & Deininger 1991; Novick et al. 1993). Some Alu insertion polymorphisms have been identified by chance (Deininger & Batzer 1995) while others have been identified by library screening in a directed approach (Batzer & Deininger 1991; Batzer et al. 1995; Arcot et al. 1995a; Arcot et al. 1995b; Arcot et al. 1995c; Batzer et al. 1996a; Arcot et al. 1998). Here, we have presented two complementary methods involving computational biology and PCR based displays that will enhance our ability to identify the genomic fossils of recently integrated mobile elements from complex genomes. These approaches represent the beginning of a new era in biological sciences that will increasingly rely upon informatics/computational biology as well as hard-core bench molecular biology to answer global questions in comparative genomics.

ACKNOWLEDGEMENTS

DHK was supported by an Eastern Michigan University Spring/Summer Research Award and a University Research in Excellence Fund. AS was supported by a fellowship from the Academy of Scientific Research, Ministry of Scientific Research and Informatics, Cairo, Egypt. This research was supported by National Institutes of Health RO1 GM45668 to PLD, Department of the Army DAMD17-98-1-8119 to PLD and

MAB, and award number 1999-IJ-CX-K009 from the Office of Justice Programs, National Institute of Justice, Department of Justice MAB. Points of view in this document are those of the authors and do not necessarily represent the official position of the U.S. Department of Justice.

REFERENCES

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, & D.J. Lipman, 1990. Basic local alignment search tool. J.Mol.Biol. 215:403-410.

Arcot, S.S., T. H. Shaikh, J. Kim, L. Bennett, M. Alegria-Hartman, D. O. Nelson, P. L. Deininger, & M. A. Batzer, 1995a. Sequence diversity and chromosomal distribution of 'young' Alu repeats. Gene 163:273-8.

Arcot, S.S., Z. Wang, J. L. Weber, P. L. Deininger, & M. A. Batzer, 1995b. *Alu* repeats: A source for the genesis of primate microsatellites. Genomics 29:136-144.

Arcot, S.S., A.W. Adamson, G.W. Risch, J. LaFleur, M.B. Robichaux, J.E. Lamerdin, A.V. Carrano, & M.A. Batzer, 1998. High-resolution cartography of recently integrated human chromosome 19- specific Alu fossils. J.Mol.Biol. 281:843-856.

Arcot, S.S., J.J. Fontius, P.L. Deininger, & M.A. Batzer, 1995c. Identification and analysis of a 'young' polymorphic *Alu* element. Biochem.Biophys.Acta 1263:99-102.

Ausubel, F.M., R. Brent, R.E. Kingston, D.D. Moore, J.G. Seidman, J.A. Smith, & K. Struhl, 1996. Current Protocols In Molecular Biology. John Wiley & Sons, Inc., Canada.

Batzer, M.A., P.L. Deininger, U. Hellmann-Blumberg, J. Jurka, D. Labuda, C.M. Rubin, C.W. Schmid, E. Zietkiewicz, & E. Zuckerkandl, 1996b. Standardized nomenclature for *Alu* repeats. J.Mol.Evol. 42:3-6.

Batzer, M.A., C.M. Rubin, U. Hellmann-Blumberg, M. Alegria-Hartman, E.P. Leeflang, J.D. Stern, H.A. Bazan, T.H. Shaikh, P.L. Deininger, & C.W. Schmid, 1995. Dispersion and insertion polymorphism in two small subfamilies of recently amplified human Alu repeats. J.Mol.Biol. 247:418-427.

Batzer, M. A., M. Alegria-Hartman, H. Bazan, D. H. Kass, G. Novick, P. A. Ioannou, D. Boudreau, W.D. Scheer, R. J. Herrera, M. Stoneking, & P. Deininger, 1994a. Alu repeats as markers for human population genetics. IVth International Symposium on Human Identification. 49-57.

Batzer, M. A., S.S. Arcot, J. W. Phinney, M. Alegria-Hartman, D. H. Kass, S. M. Milligan, C. Kimpton, P. Gill, M. Hochmeister, P. A. Ioannou, R. J. Herrera, D. A. Boudreau, W.D. Scheer, B. J. B. Keats, P. L. Deininger, & M. Stoneking, 1996a. Genetic variation of recent *Alu* insertions in human populations. J.Mol.Evol. 42:22-29.

Batzer, M. A. & P. L. Deininger, 1991. A human-specific subfamily of *Alu* sequences. Genomics 9:481-487.

Batzer, M. A., V. Gudi, J.C. Mena, D. W. Foltz, R. J. Herrera, & P. L. Deininger, 1991. Amplification dynamics of human-specific (HS) Alu family members. Nucleic Acids Res. 19:3619-3623.

Batzer, M. A., G. E. Kilroy, P. L. Richard, T. H. Shaikh, T. D. Desselle, C. L. Hoppens, & P. L. Deininger, 1990. Structure and variability of recently inserted Alu family members. Nucleic Acids Res. 18:6793-6798.

Batzer, M. A., M. Stoneking, M. Alegria-Hartman, H. Bazan, D. H. Kass, T. H. Shaikh, G. Novick, & P.A. Ioannou, 1994b. African origin of human-specific polymorphic Alu insertions. Proc.Natl.Acad.Sci.,USA 91:12288-12292.

Batzer, M.A., C.W. Schmid, & P.L. Deininger, 1993. Evolutionary analyses of repetitive DNA sequences. Methods Enzymol. 224:213-32:213-232.

Bird, A.P., 1980. DNA methylation and the frequency of CpG in animal DNA. Nucleic.Acids.Res. 8:1499-1504.

Daniels, G. & P. L. Deininger, 1985. Repeat sequence families derived from mammalian tRNA Genes. Nature 317:819-822.

Deininger, P. L. & M. A. Batzer, 1995. SINE master genes and population biology. In: The Impact of Short, Interspersed Elements (SINEs) on the Host Genome. R.Maraia, ed. R.G. Landes, Georgetown, TX, pp. 43-60.

Deininger, P. L. & M.A. Batzer 1993 Evolution of Retroposons. In: Evolutionary Biology. M.K. Heckht and et al, eds. Plenum Publishing, New York, pp. 157-196.

Deininger, P.L. & M.A. Batzer, 1999. Alu repeats and human disease. Mol.Genet.Metab. 67:183-193.

Hammer, M.F., 1994. A recent insertion of an Alu element on the Y chromosome is a useful marker for human population studies. Mol.Biol.Evol. 11:749-761.

Hutchinson, G.B., S.E. Andrew, H. McDonald, Y.P. Goldberg, R. Graham, J.M. Rommens, & M.R. Hayden, 1993. An Alu element retroposition in two families with Huntington disease defines a new active Alu subfamily. Nucleic.Acids.Res. 21:3379-3383.

Kass, D. H., C. Alemán, M. A. Batzer, & P. L. Deininger, 1994. An HS Alu insertion caused a factor XIIIB gene RFLP. Genetica, 94:1-8.

Kass, D.H., M.A. Batzer, & P.L. Deininger, 1996. Characterization and population diversity of interspersed repeat sequence variants (IRS-morphs). Genome 39:688-696.

Labuda, D. & G. Striker, 1989. Sequence conservation in Alu evolution. Nucleic.Acids.Res. 17:2477-2491.

Miyamoto, M.M., J.L. Slightom, & M. Goodman, 1987. Phylogenetic relations of humans and African apes from DNA sequences in the psi eta-globin region. Science 238:369-373.

Munroe, D.J., M. Haas, E. Bric, T. Whitton, H. Aburatani, K. Hunter, D. Ward, & D.E. Housman, 1994. IRE-bubble PCR: a rapid method for efficient and representative amplification of human genomic DNA sequences from complex sources. Genomics 19:506-514.

Novick, G., T. Gonzalez, J. Garrison, C. Novick, M. Batzer, P. Deininger, & R.Herrera, 1993. The use of polymorphic Alu insertions in human DNA fingerprinting. In: DNA Fingerprinting: State of the Science. S.D.J.Pena, R.Chakraborty, J.T.Epplen and A.J.Jeffreys, eds. Birkhauser Verlag, Basel, pp. 283-291.

Perna, N. T., M. A. Batzer, P. L. Deininger, & M. Stoneking, 1992. Alu insertion polymorphism: A new type of marker for human population studies. Human Biology 64:641-648.

Shen, M. R., M . A. Batzer, & P. L. Deininger, 1991. Evolution of the Master Alu Gene(s). J.Mol.Evol. 33:311-320.

Stoneking, M., J.J. Fontius, S.L. Clifford, H. Soodyall, S.S. Arcot, N. Saha, T. Jenkins, M.A. Tahir, P.L. Deininger, & M.A. Batzer, 1997. Alu insertion polymorphisms and human evolution: evidence for a larger population size in Africa. Genome Res. 7:1061-1071.

Zietkiewicz, E., C. Richer, W. Makalowski, J. Jurka, & D. Labuda, 1994. A young Alu subfamily amplified independently in human and African great apes lineages. Nucleic.Acids.Res. 22:5608-5612.

FIGURE LEGENDS

Figure 1. Computational analysis of repetitive elements. The flow chart shows the computational tools utilized for the identification and analysis of recently integrated Ya8 Alu family members. The process begins with BLAST searches of the non-redundant and high-throughput genomic sequence databases. Subsequently sequences (about 1000 nucleotides) adjacent to the matches with 100% identity to the query sequence are annotated using the RepeatMasker2 server. Following sequence annotation, oligonucleotide primers complementary to the unique DNA sequences adjacent to each element are designed using the Primer3 web server. The oligonucleotide designed using Primer3 are then subjected to a second BLAST search to determine if they reside in other repetitive elements and they are then used for PCR based analyses of individual mobile elements.

Figure 2. Multiple alignment of Ya8 subfamily members. The Ya8 subfamily consensus (con) is derived from the most common nucleotide found at each position within the subfamily members. Nucleotide substitutions at each position are indicated with the appropriate nucleotide. Deletions are marked by an -.

Figure 3. Nucleotide sequences flanking Ya8 subfamily members. Nucleotide sequences flanking the Ya8 Alu family members are shown. Nucleotides encompassed in the direct repeats are underlined. The length of the oligo-dA rich tail is denoted by an (A) and a subscript indicating the number of adenine residues.

Figure 4. The Allele-Specific Alu PCR (ASAP) anchor strategy. Schematic diagram of the technique for the isolation of a designated subset of Alu repeats based on a modification of the IRE-bubble PCR technique (Munroe et al. 1994). The shaded rectangle represents an Alu sequence in genomic DNA. The *Mse*I cleaves in unique sequences flanking the Alu repeat (small arrows). The anchors with the complementary *Mse*I site are ligated. The anchors are designed so that the two oligonucleotide stands base-pair only at the *Mse*I site end, but not at the other end (represented schematically with four bases). PCR is initiated using an allele-specific Alu primer (Z'). The anchor primer will not be able to base pair preventing anchor-to-anchor amplification. Only those fragments (a) generated by the Alu primer are available for amplification by the anchor primer. The amplified product (a and b) provides a template for nested PCR (primer y') to further decrease the background.

Figure 5. DNA fingerprints of unrelated individuals based on anchored-Alu PCR. Individual bands are numbered for identification purposes. Molecular weights are shown in nucleotides to the left. DNA samples used are of Caucasian (lane a), Hispanic (lane b), Hindu-Indian (lane c) and Chinese (lane d) descent.

Table 1: Ya8 Accession Numbers, Primers, Location, and Product

Name	Accession #	5' Primer sequence (5'-3')	3' Primer sequence (5'-3')	A.T.	Chromosomal Location	Product Size	
						Filled	Empty
Ya8NBC1	AC006959	CCTGCTGACATTTAGAAATGACTCT	ATATACAAGTCATCAGATGGGGACAC	60°C	5	504	293
Ya8NBC2	AC006556	GCCTGTGTACCTCCTTTAAATATCTTG	CTCAAAAACTGGAGCAGGAGTAA	50°C	21	503	242
Ya8NBC3*	AC006989	GGTGGTCATCCATATACTATCTCATAGG	AGAGTTCTGGAAAAGTTGACAGGAT	55°C	Y/X	498	178
Ya8NBC4	AL049871	CATTCCACCCTGTCAGCATT	GCTTTGGAAGTAGGCAGGTTAC	60°C	14	536	204
Ya8NBC6	AC004066	ACTTAGCTTTGAGTATTTTTCTGAACTATC	CTAAAATGGAGGTACCGATATACTTTATTA	60°C	4	470	132
Ya8NBC8	AL034422	GGATCACAAACCTAAATGAAAGAGGTAA	CCGTCTCAAAACAAACAGACAAATA	60°C	20	501	155
Ya8NBC10	AC004893	ATAGTGGTCTTCAAAGTACAATCCAGTT	TTTTCAGCAAAACCCTAAACCTAGT	60°C	7	507	200
Ya8NBC11	AC007688	GAGTGCCTATTATGTGTTAGGTACTTTGCT	ACTCTCACTAGATTATAAGCCCCATAAGGA	60°C	12	419	105
Ya8NBC12	AL022302	CATCTTAAAAGACATTAGAAAAGTACACAG	CTGGCCACTTAGTATATTTTCAATCAG	60°C	22	530	211
Ya8NBC13	AL008722	CCATTTTCTATAAGAAGGCTTCACC	AAAGTAATGTGAAAGTATTGGAGAAGAGAT	60°C	22	402	77
Ya8NBC14	AF094481	GAATCTCTATCTCTGACACTAGCCACT	GGCAACAAGTCTGATGAATACTTAAAGGAG	60°C	3	500	189
Ya8NBC15	AF179296	CTCTACAGTACAGATGAGAAAGTACAGACA	CGCCTTGCTAGATTTCTTTCTAATG	60°C	8	620	299
Ya8NBC17	AC005205	CTAGTTCCCACATACCGAAAACAC	CCTGTCTCGTTCAGTCTTCTTTG	58°C	19	501	155
Ya5NBC60	AC006553	CAGTCCATAGCAGTCATGGTAAATAAG	AAGTCTATACCGGTTACCTCTTTCTT	58°C	4	456	149

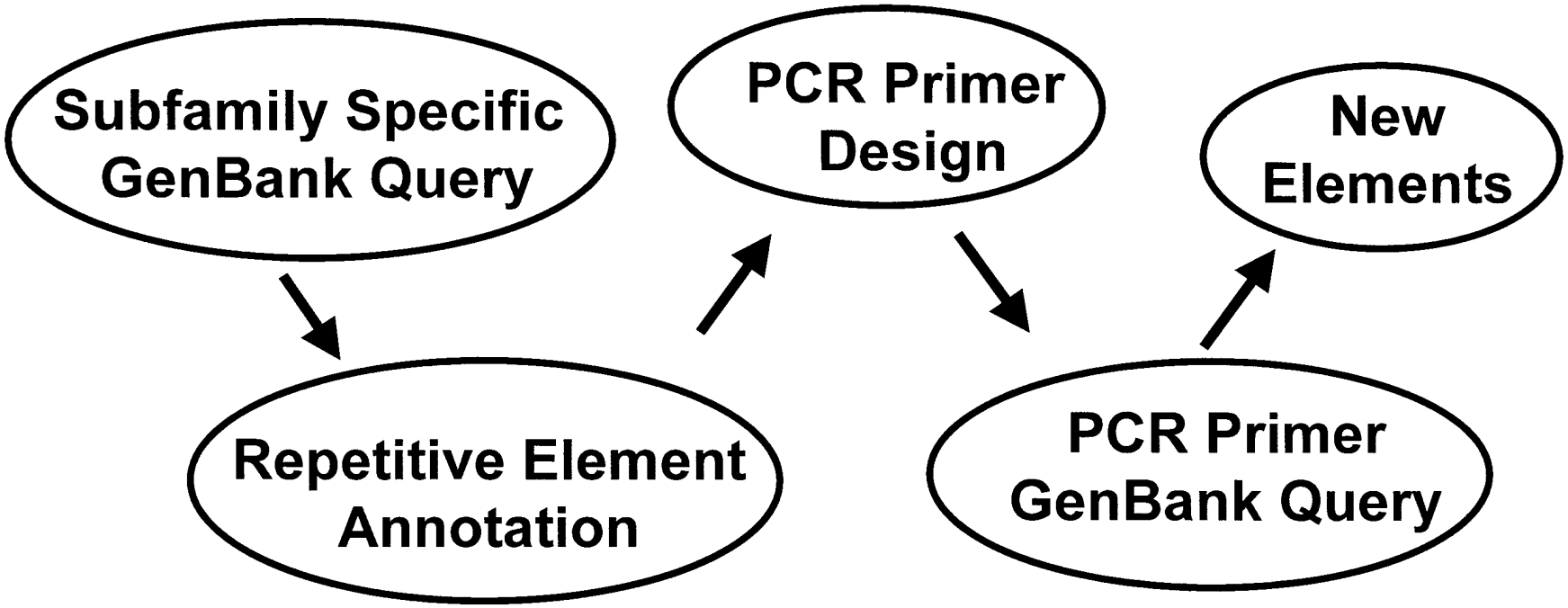
1. Amplification of each locus required 2:30 min @ 94°C initial denaturing, and 32 cycles for 1min 94°C, 1 min Annealing Temperature (A.T.) and 1 min elongation at 72°C. Extension time of 10 min at 72°C was used.
3. Chromosomal location determined from Accession information or by PCR analysis of monochromosomal hybrid cell lines.
4. Empty product sizes calculated by removing the Alu element and 1 direct repeat from the filled sites that were identified.
5. Ya8NBC3 is located in the pseudoautosomal region of the X and Y chromosome.

Table 2: Alu Ya8 Associated Human Genomic Diversity

Elements	African American					European					Greenland Natives					Egyptian					
	Genotypes			ƒAlu	Het	Genotypes			ƒAlu	Het	Genotypes			ƒAlu	Het	Genotypes			ƒAlu	Het[1]	Avg Het[2]
	+/+	+/-	-/-			+/+	+/-	-/-			+/+	+/-	-/-			+/+	+/-	-/-			
Ya8NBC1	10	2	7	0.58	0.50	5	0	9	0.36	0.35	10	5	1	0.78	0.48	8	0	10	0.44	0.51	0.46
Ya8NBC3	0	12	7	0.32	0.44	0	6	14	0.15	0.44	0	12	7	0.32	0.26	0	9	10	0.24	0.51	0.41
Ya8NBC4	1	4	13	0.17	0.29	6	0	7	0.46	0.51	8	5	6	0.55	0.52	18	0	1	0.95	0.10	0.35
Ya8NBC6	8	2	6	0.56	0.51	11	0	3	0.85	0.00	16	0	0	1.00	0.35	12	2	3	0.76	0.37	0.31
Ya8NBC11	13	2	0	0.93	0.13	12	0	0	1.00	0.09	10	1	0	0.95	0.00	13	3	0	0.91	0.18	0.10
Ya8NBC12	17	0	0	1.00	0.00	19	0	0	1.00	0.05	18	1	0	0.97	0.00	17	0	0	1.00	0.00	0.01
Ya8NBC60	6	9	3	0.58	0.50	6	7	5	0.53	0.51	5	9	3	0.56	0.51	10	5	4	0.66	0.46	0.49

1. This is the unbiased heterozygosity.
2. Average heterozygosity is the average of the population heterozygosity.

Figure 1


Subfamily Specific GenBank Query
PCR Primer Design
New Elements
Repetitive Element Annotation
PCR Primer GenBank Query

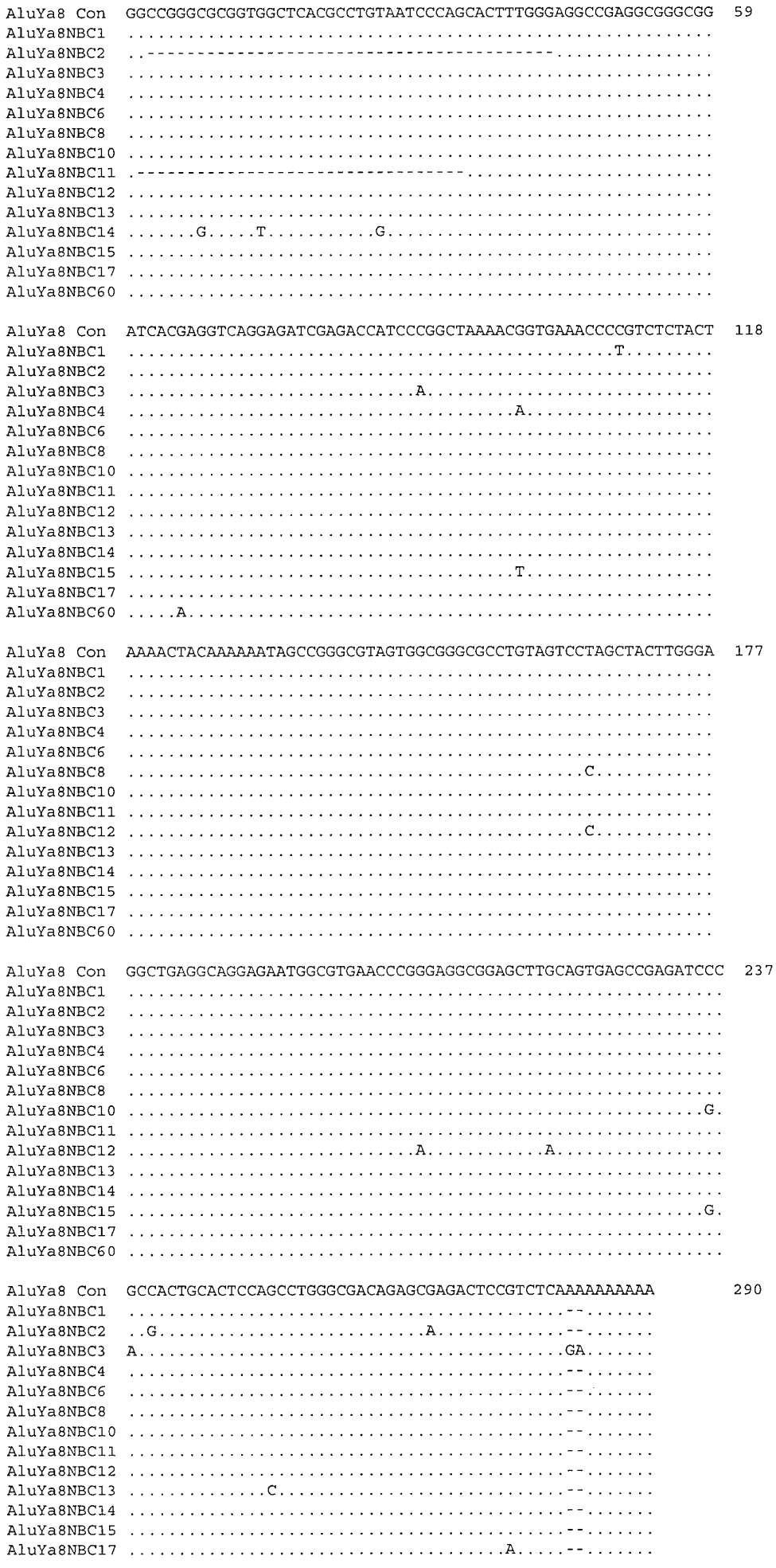

AluYa8 Con GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGCGG 59
AluYa8NBC1 .....
AluYa8NBC2 ..----------.....
AluYa8NBC3 .....
AluYa8NBC4 .....
AluYa8NBC6 .....
AluYa8NBC8 .....
AluYa8NBC10 .....
AluYa8NBC11 .----------.....
AluYa8NBC12 .....
AluYa8NBC13 .....
AluYa8NBC14 .....G.....T.....G.....
AluYa8NBC15 .....
AluYa8NBC17 .....
AluYa8NBC60 .....
AluYa8 Con ATCACGAGGTCAGGAGATCGAGACCATCCCGGCTAAAACGGTGAAACCCCGTCTCTACT 118
AluYa8NBC1 .....T.....
AluYa8NBC2 .....
AluYa8NBC3 .....A.....
AluYa8NBC4 .....A.....
AluYa8NBC6 .....
AluYa8NBC8 .....
AluYa8NBC10 .....
AluYa8NBC11 .....
AluYa8NBC12 .....
AluYa8NBC13 .....
AluYa8NBC14 .....
AluYa8NBC15 .....T.....
AluYa8NBC17 .....
AluYa8NBC60 .....A.....
AluYa8 Con AAAACTACAAAAAATAGCCGGGCGTAGTGGCGGGCGCCTGTAGTCCTAGCTACTTGGGA 177
AluYa8NBC1 .....
AluYa8NBC2 .....
AluYa8NBC3 .....
AluYa8NBC4 .....
AluYa8NBC6 .....
AluYa8NBC8 .....C.....
AluYa8NBC10 .....
AluYa8NBC11 .....
AluYa8NBC12 .....C.....
AluYa8NBC13 .....
AluYa8NBC14 .....
AluYa8NBC15 .....
AluYa8NBC17 .....
AluYa8NBC60 .....
AluYa8 Con GGCTGAGGCAGGAGAATGGCGTGAACCCGGGAGGCGGAGCTTGCAGTGAGCCGAGATCCC 237
AluYa8NBC1 .....
AluYa8NBC2 .....
AluYa8NBC3 .....
AluYa8NBC4 .....
AluYa8NBC6 .....
AluYa8NBC8 .....
AluYa8NBC10 .....G.
AluYa8NBC11 .....
AluYa8NBC12 .....A.....A.....
AluYa8NBC13 .....
AluYa8NBC14 .....
AluYa8NBC15 .....G.
AluYa8NBC17 .....
AluYa8NBC60 .....
AluYa8 Con GCCACTGCACTCCAGCCTGGGCGACAGAGCGAGACTCCGTCTCAAAAAAAAAA 290
AluYa8NBC1 .....--.....
AluYa8NBC2 ..G.....A.....--.....
AluYa8NBC3 A.....GA.....
AluYa8NBC4 .....--.....
AluYa8NBC6 .....--.....
AluYa8NBC8 .....--.....
AluYa8NBC10 .....--.....
AluYa8NBC11 .....--.....
AluYa8NBC12 .....--.....
AluYa8NBC13 .....C.....--.....
AluYa8NBC14 .....--.....
AluYa8NBC15 .....--.....
AluYa8NBC17 .....A.....--.....

Figure 2

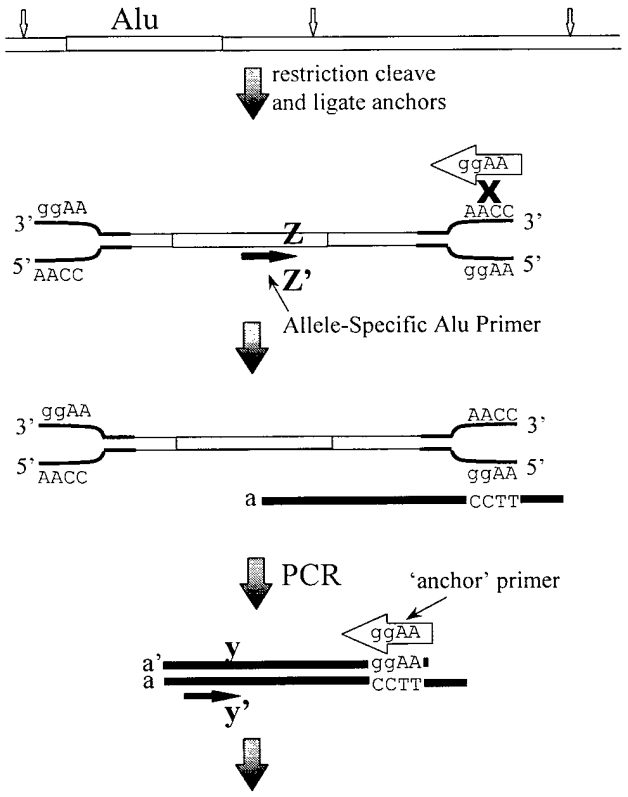
AluYa8NBC60 .....--.....

Figure 2 Continued

Ya8NBC1	AAGAGGGGGAGAG	[Alu]	A_{18} AAGAGGGGGAGAG
Ya8NBC2	GGA	[Alu]	$A_{16}C(A)_4$ TGGA
Ya8NBC3	GAAGAAGTTTTGC	[Alu]	$AC(A)_{21}C(A)_2$ GAAGAAGTTTTGC
Ya8NBC4	CGACAATTT	[Alu]	$A_{17}C(A)_{13}C(A)_{10}$ CCGACAATTT
Ya8NBC6	AAATTTAAAATATT	[Alu]	A_{44} AAATTTAAAATATT
Ya8NBC8	AAGAAAATATAGGCATA	[Alu]	$A_{11}C(A)_{14}C(A)_{23}$ AAGAAAATATAGGCATA
Ya8NBC10	AATGAATTTTTTAGG	[Alu]	A_{12} AATGCATTTTTTAGG
Ya8NBC11	AAGGAATGAGACTG	[Alu]	A_{20} AAGGAATGAGACTG
Ya8NBC12	AAAGTTCTTTGCA	[Alu]	A_{27} AAAGTTCTTTGCA
Ya8NBC13	AAGAAGGCTTCACCAG	[Alu]	A_{30} AAGAAGGCTTCACCAG
Ya8NBC14	ATCCC	[Alu]	A_{26} ATCCC
Ya8NBC15	AGAACCACCAGGAA	[Alu]	A_{27} AGAACCACCAGGAA
Ya8NBC17	AAGGAATCTC	[Alu]	A_{17} AAGGAATCTC
Ya8NBC60	GGTAAATAAGCTTTCTT	[Alu]	A_{25} GGTAAATAAGCTTTCTT

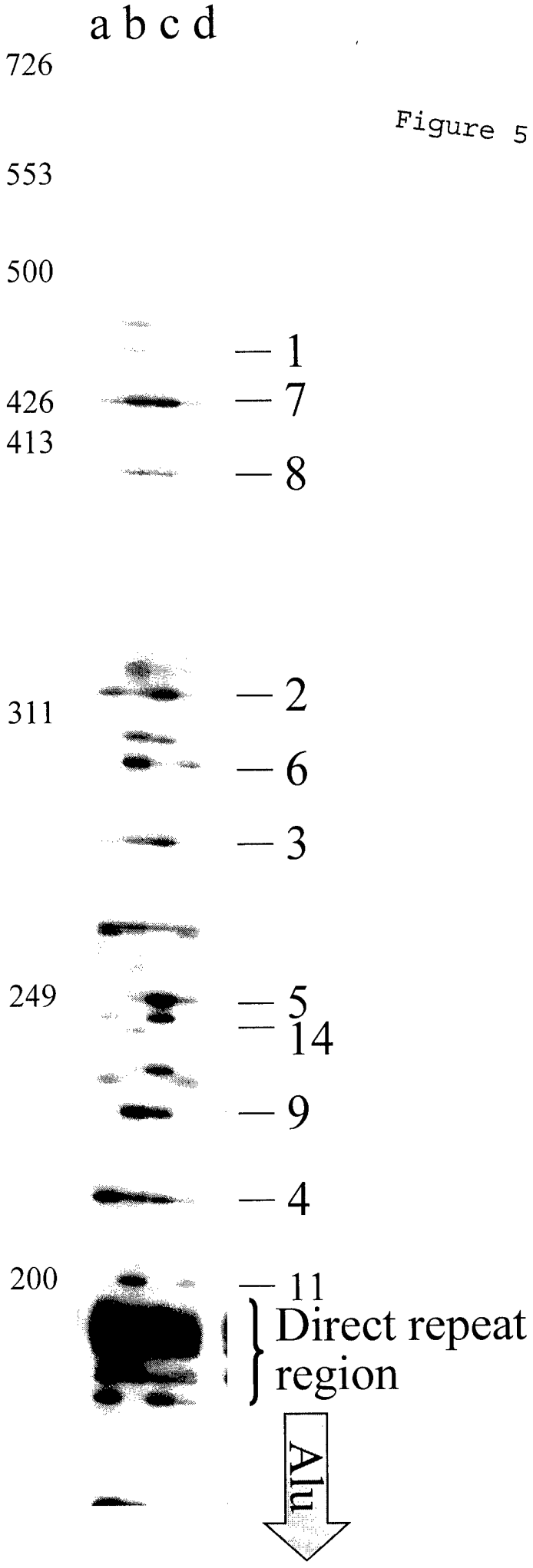
Figure 3

Figure 4



Nested Allele-Specific Alu PCR

Figure 5


a b c d
726
553
500
426
413
311
249
200
1
7
8
2
6
3
5
14
9
4
11
Direct repeat region
Alu

Potential gene conversion and source gene(s) for recently integrated Alu elements

Running title: Mosaic Alu sequences

Keywords: Alu insertion polymorphism, gene conversion, SINE, source gene

Abbreviations: FGFR2- fibroblast growth-factor receptor 2; NF-1- neurofibromatosis 1; IL2RG- interleukin 2 receptor gene; BRCA2- breast cancer 2.

Astrid M. Roy[1,+], Marion L. Carroll[2,+], Son V. Nguyen[2], Abdel-Halim Salem[2], Michael Oldridge[3], Andrew O. M. Wilkie[3,4], Mark A. Batzer[2,^], and Prescott L. Deininger[1,5,^ *]

[1]Tulane Cancer Center, Department of Environmental Health Sciences, Tulane University - Medical Center, 1430 Tulane Ave., SL-66, New Orleans, Louisiana 70112.
[2]Departments of Pathology, Biometry and Genetics, Biochemistry and Molecular Biology, Stanley S. Scott Cancer Center, Neuroscience Center of Excellence, Louisiana State University Health Sciences Center, 1901 Perdido Street, New Orleans, Louisiana 70112.
[3]Institute of Molecular Medicine, John Radcliffe Hospital, Oxford, UK.
[4]Oxford Craniofacial Unit, The Radcliffe Infirmary NHS Trust, Oxford, UK.
[5]Laboratory of Molecular Genetics, Alton Ochsner Medical Foundation, 1516 Jefferson Highway, New Orleans, Louisiana 70121.

+AMR and MLC contributed equally to this research.

^MAB and PLD are equal senior authors.

* Corresponding author.
Mailing address: Tulane Cancer Center
Tulane University - Medical Center
1430 Tulane Ave., SL-66
New Orleans, LA 70112
Phone: (504) 988- 6385
Fax: (504) 588-5516
Internet: PDEININ@TCS.TULANE.EDU

ABSTRACT

Alu elements comprise greater than 10% of the human genome. We have employed a computational biology approach to analyze the human genomic DNA sequence databases to determine the impact of gene conversion on the sequence diversity of recently integrated Alu elements, and to identify Alu elements that were potentially retroposition competent. We analyzed 269 Alu Ya5 elements and identified 23 members of a new Alu subfamily termed Ya5a2 with an estimated copy number of 35 members, including the *de novo* Alu insertion in the NF1 gene. Our analysis of Alu elements containing one to four (Ya1-Ya4) of the Ya5 subfamily-specific mutations suggests that gene conversion contributed as much as 10-20% to the variation between recently integrated Alu elements. In addition, analysis of the middle A-rich region of the different Alu Ya5 members indicates a tendency toward expansion of this region and subsequent generation of simple sequence repeats. Mining the databases for putative retroposition competent elements that share 100% nucleotide identity to the previously reported *de novo* Alu insertions linked to human diseases resulted in the retrieval of 13 exact matches to the NF1 Alu repeat, three to the Alu element in BRCA2, and one to the Alu element in FGFR2 (Apert syndrome). Transient transfections of the potential source gene for the Apert's Alu with its endogenous flanking genomic sequences demonstrated the transcriptional and presumptive transpositional competency of the element.

INTRODUCTION

Alu elements belong to a class of retroposons termed SINEs. SINEs are Short INterspersed Elements usually about 100 – 300 bp in length commonly found in introns, 3' untranslated regions of genes, and intergenic genomic regions (Deininger and Batzer 1993). Alu is the most abundant class of SINEs in primate genomes, reaching a copy number in excess of one million/haploid genome (Jelinek and Schmid 1982; Jurka et al. 1993, Smit 1999). Alu elements increase their genomic copy number by an amplification process termed retroposition (Rogers and Willison 1983; Weiner et al. 1986).

Alu elements appear to have arisen in the last 65 million years (my) (Deininger and Daniels 1986). The human Alu family of repeats is composed of a small number of distinct subfamilies characterized by subfamily-specific diagnostic mutations (Batzer et al. 1996; Slagel et al. 1987; Willard et al. 1987; Shen et al. 1991). The source Alu gene(s) for each of the subfamilies has been retropositionally active during different periods of primate evolution. The rate of Alu amplification (mostly Sx subfamily) appears to have reached its peak between 60 and 35 my, and subsequently decreased several orders of magnitude to the present amplification rate (Shen et al. 1991). Only a limited number of SINEs, termed "master" or source genes, appear to be capable of retroposition (Batzer et al. 1990; Deininger et al. 1992; Deininger and Daniels 1986), although the critical factor(s) defining functional source genes are not understood. A variety of factors influence the retroposition process (Schmid and Maraia 1992). Currently, only the recently integrated "young" Alu subfamilies appear to be retropositionally active. Almost all of the recently integrated Alu elements within the human genome belong to one of four closely related subfamilies: Y, Ya5, Ya8, and Yb8,

with the majority being Ya5 and Yb8 subfamily members (Batzer et al. 1990; Batzer et al. 1995; Deininger and Batzer, 1999).

Previously, analysis of individual Alu elements from the different subfamilies involved laborious procedures, such as cloning, library screening and subsequent sequencing (Batzer et al. 1990; Batzer et al. 1995; Arcot et al. 1995a). However, the availability of large-scale human genomic DNA sequences as a result of the Human Genome Project facilitates genomic database mining for Alu elements (Roy et al. 2000). We have taken advantage of these databases and analyzed a significant portion of the Alu Ya5 subfamily, as well as intermediates between the Ya5 subfamily and the ancestral Alu Y subfamily. In addition, we searched the databases for putative retroposition competent source Alu genes that generated the *de novo* Alu inserts associated with a number of human diseases (Deininger and Batzer 1999).

RESULTS

Computational Analyses

In order to search for previously unidentified subfamilies within the Ya5 Alu subfamily, we selected all of the Alu family members that matched our Ya5 consensus query sequence from the human genome non-redundant (nr) database. Only Ya5 elements randomly found within other sequences were included in our analysis. This eliminated Alu elements that had been previously identified in directed Alu–specific projects. In addition, truncated Alu elements were eliminated from the analysis. Ya4 elements that did not contain the first Ya5 specific diagnostic mutation #11 (Fig. 1) (Shen et al. 1991), which is a CpG dinucleotide in the Ya5 subfamily, were considered as Ya5 Alu family members. We obtained a total of 269 matches to the Ya5 query sequence that met our criteria. Of these, 47 shared 100% nucleotide identity with the subfamily consensus sequence and 83 were near perfect matches aside from a few CpG mutations.

Analysis of the 269 Ya5 Alu elements resulted in the initial identification of two subsets of potential "subfamilies" containing two diagnostic mutations each, one with six members and the other with four. They will be referred to as Ya5a2 and Ya5b2 respectively, in compliance with the standard Alu subfamily nomenclature (Batzer et al. 1996). Each consensus sequence with the two diagnostic mutations specific to each new Alu subfamily are shown in Figure 1. Interestingly, the *de novo* Alu Ya5 insert present within an intron of the NF1 gene (Wallace et al. 1991) is an exact match to the Ya5a2 consensus. The nr database contained 16.0% of human DNA sequences for a total of 515,596,000 bases on the date of the search. The estimated size of the Ya5a2 subfamily is (3×10^9 bp/ 515,596,000 bp) x 6 unique Ya5a2 matches = 35 subfamily members. In

comparison, the estimated size of the Ya5b2 subfamily is (3 x 10^9 bp/ 515,596,000 bp) x 4 unique Ya5b2 matches = 22 subfamily members. We utilized only the randomly found Ya5a2 elements for the calculations to avoid overestimating the size of the subfamilies. However, these numbers may be underestimations, because some specific polymorphic elements of this subfamilies may not be represented in the database.

In order to derive a second estimate of the copy numbers of the Ya5a2 and Ya5b2 Alu subfamilies, we used their consensus sequences as queries for the high throughput genome sequence (htgs) and genomic survey sequence (gss) databases. Seventeen additional Alu Ya5a2 elements were found in these searches. Of the 23 total Ya5a2 elements, 13 shared 100% nucleotide identity with the subfamily consensus sequence. No additional Ya5b2 elements were found in the other databases, therefore the Ya5b2 subfamily was not subjected to further analysis. Three additional potential subfamilies, Ya5a1 (5 members), Ya5b1 (4 members), and Ya5c1 (4 members) with only one specific diagnostic mutation were identified (Figure 1). Due to the small copy number, and the possibility that some of those represent parallel mutations rather than subfamilies, no further analyses were performed.

To determine the age of the Ya5a2 subfamily, we divided the nucleotide substitutions within the elements into those that have occurred in CpG dinucleotides and those that have occurred in non-CpG positions. The distinction between types of mutations is made because the CpG dinucleotides mutate at a rate that is about 10 times faster than non-CpG (Labuda and Striker 1989; Batzer et al. 1990), as a result of the deamination of 5-methylcytosine (Bird 1980). A total of five non-CpG mutations and seven CpG mutations occurred within the 23 Alu Ya5a2 subfamily members identified.

Using a neutral rate of evolution for primate intervening DNA sequences of 0.15% per million years (Miyamoto et al. 1987) and the non-CpG mutation rate of 0.092% (5/5382 bases using only non-CpG bases) within the 23 Ya5a2 Alu elements yield an estimated average age of 0.62 million years (my) for the Ya5a2 subfamily members with a predicted 95% confidence level in the range of 0.28-1.08 my given that the mutations were random and fit a binomial distribution. The Ya5a2 subfamily appears to be much younger than Ya5, Ya8, or Yb8 Alu subfamilies with estimated ages of 2.8 my (Batzer et al. 1990), 2.75 my (Roy et al. 2000) and 2.7 my (Batzer et al. 1995), respectively (Fig. 2).

Determination of the number of elements that perfectly match the subfamily consensus sequence can also give an indirect estimate of Alu subfamily age and recent rate of mobilization. Recently transposed Alu elements share higher levels of nucleotide identity with their source copies since they have not resided in the genome long enough to accumulate random mutations. By contrast, older Alu elements that have resided in the genome for longer periods of time tend to have less nucleotide identity with their source genes as a result of the accumulation of random mutations subsequent to integration into the genome. We compared our search results for the Ya5a2 subfamily with parallel searches from the Ya8, and Ya5 Alu subfamilies. Our BLAST searches from the nr database yielded one perfect match out of 12 elements for Ya8, 47 out of 269 for Ya5, and 3 out of 6 for Ya5a2 (Fig. 2). Searching all three databases (nr, gss and htgs) yielded 5 perfect matches out of 27 for Ya8 and 13 out of 23 for Ya5a2. These results are in good agreement with the previous estimates indicating Ya5a2 is the

youngest Alu subfamily reported to date since it also has the highest proportion of elements that share 100% nucleotide identity with the consensus sequence.

Stability of the middle A-rich region in Alu Ya5 members

The oligo-dA rich tails and middle A-rich regions of Alu elements have previously been shown to serve as nuclei for the genesis of simple sequence repeats (Arcot et al. 1995b). In the autosomal recessive neurodegenerative disease, Friedreich ataxia, the most common mutation is the hyperexpansion of a GAA within the middle A-rich region of an Sx Alu element (Montermini et al. 1997). Since these regions appear unstable, we analyzed the middle A-rich region of Alu elements retrieved from the databases to detect expansions/contractions of this sequence.

To evaluate potential expansions/contractions we performed a BLAST query of three databases (nr, htgs, and gss) using the Alu Ya5 consensus sequence with varying numbers of A nucleotides within the middle A rich-region (TA_nTACA_nTT). Our results demonstrate that the majority of the elements identified matched the consensus sequence. However, there is a trend for an A expansion at both positions (Table 1). By contrast, very few sequence contractions were detected for any of the positions.

Human genomic variation

In order to determine the human genomic variation associated with the Ya5a2 Alu subfamily members, we selected the 13 Ya5a2 elements identical to the subfamily consensus sequence as well as two others and determined the degree of fixation associated with the elements using PCR based assays of a panel of diverse human DNA

samples using the primers shown in Table 2. The panel is composed of 20 individuals of European origin, African-Americans, Greenland Natives, and Egyptians for a total of 80 individuals (160 chromosomes). The Alu elements were classified as fixed absent, fixed present and high, intermediate or low frequency insertion polymorphisms (see Table 3 for definitions). Using this approach three of the 14 elements tested (Ya5NBC206, 207, and 235) were always present in the human genomes that were surveyed, suggesting that these elements became fixed in the genome prior to the radiation of modern humans from Africa. Six of the elements (Ya5NBC208, 236, 240, 241, 242, and 220) are intermediate frequency Alu insertion polymorphisms. The remaining six elements are low frequency Alu insertion polymorphisms (Table 3). The population specific genotypes and levels of heterozygosity for each element are shown in Table 4. The high proportion of polymorphic elements is in good agreement with our previous observations indicating that the Ya5a2 subfamily is younger than any of the other Alu subfamilies previously identified in the human genome.

Gene conversion and Alu sequence diversity

In our query of the human genome (nr) database, 91 of the Alu elements identified contain one to four of the five Ya5 diagnostic nucleotides (Fig. 1). Of these 91 "intermediate" elements, four are Ya1, one Ya2, seven Ya3, and 79 Ya4 Alu elements (Fig. 3). Surprisingly not all of the Alu elements with different numbers of subfamily mutations had the same combination of mutations. To facilitate identification of the individual elements with different diagnostic mutation combinations, the diagnostic nucleotides were numbered consecutively in order of abundance (Ya3.1, Ya3.2 , etc., see

Fig. 3). Seventeen Alu elements (Ya4.4) did not contain the first diagnostic mutation (#11), but were still classified as Ya5 for the analyses outlined above.

Previous evolutionary analyses of the Ya5 founder element using different primate DNA samples demonstrated the sequential accumulation of the Ya5 diagnostic mutations with diagnostic positions #13/#14 first, followed by #12/#16, and finally position #11 (Shaikh and Deininger 1996). Our data are not consistent with a sequential order in the accumulation of the diagnostic mutations. The elements classified as Ya1, Ya2, Ya3.4, Ya3.5 and Ya4.4 (26 total) fit the proposed order (Fig. 3). However, the remaining 65 elements represent almost every other permutated order. Several mechanisms could explain the occurrence for mosaic Alu elements, which are addressed in the discussion section. However, we believe the most likely explanation for the existence of these mosaic elements is through gene conversion events. A limited amount of gene conversion between Yb8 Alu elements has been reported previously (Kass et al. 1995; Batzer et al. 1995). In theory, gene conversion may change the sequence of all or part of any Alu element in either an evolutionarily "forward" (Ya5 subfamily in this case) or "backward" (Y subfamily) direction by changing the diagnostic mutations. In addition, double gene conversions would be extremely rare, making the direction of the gene conversion clear in some elements. We classified the 91 mosaic Alu element sequences as gene converted forward (f), backward (b) or could not be determined (-), see Fig. 3. If the Alu elements that fit the proposed sequential evolution are ignored in the analysis, all the other elements may be classified as backward gene conversion (32 total) or could not be determined (33 total), and none were clearly gene converted forward. Therefore, backward gene conversion may have contributed to between 10% and 20%

(32 to 65/269 Ya5 + [91-17] Ya1-4) of the Alu Ya5 sequence diversity. Interestingly, evaluation of the five random Ya5a2 non-CpG mutations shows that one mutation in position #13 is a backward mutation to the Y subfamily, another putative example of a reverse gene conversion.

In search of retroposition competent Alu repeats

Sixteen different Alu insertions have been linked to human diseases (Deininger and Batzer 1999). Four belong to the Alu Y subfamily, one to the Ya4 subfamily, eight to the Ya5 subfamily, and three to the Yb8 subfamily. Closer inspection of the nucleotide sequences of these Alu elements show that they have some mutations that are different from their respective subfamily consensus sequences. Since these Alu insertions are very recent in origin, they are likely to be identical to their source genes, aside from rare mutations introduced during reverse transcription of the Alu element. Therefore, sequence database queries utilizing each Alu element along with its individual mutations (away from the subfamily consensus sequence) may facilitate the identification of the source Alu element that generated the copy. This strategy is similar to that previously used in the identification of active LINE elements from the human genome (Dombroski et al. 1993).

A database query using the sequence of the individual Alu elements responsible for each disease to mine three databases (nr, htgs, and gss) identified exact complements to four of the disease associated Alu repeats. Thirteen of the identified elements were exact matches to the NF1 Alu insertion (Ya5a2 subfamily, Table 3) (Wallace et al. 1991); three were exact matches to the BRCA2 Alu element (Miki et al. 1996)

(Accession # AL121964, AL136319, and AL135778); one matched the FGFR2 Alu repeat (Oldridge et al. 1999) (Accession # AL031274); and one matched the Alu repeat in the IL2RG gene (Lester et al. 1997) (Accession # AC010888).

Potential source gene for the Ya5 insert in FGFR2

As mentioned above, our BLAST query only detected one exact match (Accession #AL031274 or Ya5NBC237) to the Ya5 Alu found in the FGFR2 gene that caused Apert syndrome. We estimated the level of human genomic variation associated with Ya5NBC237 using the same human DNA panel and determined that it was an intermediate frequency Alu insertion polymorphism (Table 4).

Mobilization competent Alu elements must be capable of transcription, the first step in the retroposition process. To evaluate Alu Ya5NBC237 as a potential source gene for the *de novo* insert in the patient with Apert syndrome, we determined its transcription capability. Constructs with the genetic loci containing the Ya5NBC237 Alu and the *de novo* Apert syndrome Alu element were made. Transcription levels from the two constructs were evaluated by northern blot analysis relative to a control plasmid where the Alu element is flanked immediately upstream by vector sequence.

Transient transfections (Fig. 4) of the constructs into rodent cell line C6 (rat glial tumor) were performed. Although the Alu element in the control plasmid has an intact internal pol III promoter, Alu transcripts are barely detectable from the control plasmid. By contrast, the transcription from the Apert's Alu element and its potential source gene were elevated 3-4 fold, as expected for putative mobilization competent Alu repeats. This suggests that the genomic flanking sequence of Ya5NBC237 probably makes the

Alu transcription competent, one of the several requirements of a source gene. The same results were obtained from transfections in the human embryonic kidney cell line 293 (data not shown).

DISCUSSION

Our computational and experimental analyses of the Ya5 subfamily of Alu repeats provides an overall picture of the most active of the recently integrated "young" Alu subfamilies from the human genome. The analysis of Alu Ya5 repeats allowed us to address a number of questions about the biology of these elements, such as the potential impact of gene conversion events, and the identification of Alu family members from the human genome that may be capable of retroposition.

Alu elements spread throughout the genome by retroposition in the last 65 million years. The master/source gene model (Deininger et al. 1992; Shen et al. 1991; Batzer et al. 1990) posits that a very small subset of the over 1,000,000 Alu elements within the human genome are capable of high levels of retroposition; although a much larger number may make a few copies. The formation of Alu subfamilies may be explained by the sequential accumulation of mutations within the active source gene(s) followed by proliferation of the mutated source elements. A number of studies indicate that relatively few source Alu genes have played a dominant role in the amplification and evolution of Alu elements (Shen et al. 1991; Deininger and Batzer 1993; Deininger et al. 1992; Kapitonov and Jurka 1996). Although retroposition is the primary mode of SINE mobilization and sequence evolution through mutations in the source gene(s), our analysis suggests that gene conversion and genetic instability of Alu based simple sequence repeats have also had a significant impact on the sequence architecture of this major family of human genomic sequences.

There are other alternatives that could explain the occurrence of mosaic Alu elements. First, some of the mosaic Alu elements with a single mutation could be

explained by the occurrence of parallel mutations. However, this seems unlikely unless there were selection for these specific mutations, possibly through a post-transcriptional selection process (Sinnett et al. 1992). However, it is difficult to envision a selection process that would only select for mutations at adjacent diagnostic positions, as we see here. Also, recombination between different Alu elements could have generated some of these intermediate Alu elements that contain a mosaic of diagnostic mutations. However, in many cases multiple recombination events would be required to obtain this outcome, making it highly unlikely. Although there are alternative mechanisms, we believe gene conversion is the most likely explanation for the occurrence of mosaic Alu elements.

The mechanisms of genome-wide gene conversion between mobile elements are not well understood in humans (see Kass et al. 1995 and references therein). Our data show that even the very short, dispersed Alu elements are capable of high levels of gene conversion, that usually involve only short sequence stretches. In addition, our data show that reverse or backward gene conversions appear to be more favored. It seems likely that higher levels of the Y element copy number (Shen et al. 1991) or transcription (Shaikh et al. 1997) may play a role in determining the directionality of the gene conversion events. Although older Alu subfamilies, such as J and Sx are present in higher copy numbers in the genome, they diverged greatly from their consensus sequences due to mutations that have accumulated throughout evolution. Gene conversion would not be favored between such divergent sequences. However, Alu Y elements tend to be more conserved (better matches to Ya5) and with high copy number (Batzer et al. 1995). Therefore both abundance (genomic copy number and/or transcript

levels) and sequence identity appear to be influential in the Alu gene conversion events observed.

There are multiple examples of gene conversion events in literature. Genetic exchange between exogenous and different endogenous mouse L1 elements has been previously demonstrated to readily occur (Belmaaza et al. 1990). Kass et al. (1995) previously reported a gene conversion event where one of the oldest Alu family members was converted to one of the youngest Alu subfamilies Yb8. In addition, a partially converted Yb8 Alu element was also previously reported by Batzer et al., 1995. In yeast, some types of mobile elements spread through the genome by gene converting pre-existing elements (Hoff et al. 1998). When we combine this type of mobilization in the yeast genome with the Alu gene conversions reported previously, as well as those in this manuscript one could argue that gene conversion may represent a second type of amplification mechanism for short interspersed elements in the human genome. These observations suggest that evolutionary studies of all types of interspersed elements that ignore gene conversion events may lead to biased conclusions.

Variations in the length of the middle A-rich region and oligo-dA rich tails of Alu elements are not uncommon (Economou et al. 1990; Jurka and Pethiyagoda 1995; Arcot et al. 1995b). Microsatellite repeats have been found to be associated with the 3' oligo (dA) tails and the middle A-rich region of Alu elements. In the case of Friedreich ataxia the most common mutation is the hyperexpansion of a GAA trinucleotide repeat within the middle A-rich region of an Sx Alu (Montermini et al. 1997). However, microsatellites in the middle of Alu elements are not as common due to the much shorter initial length of the middle A-rich region. Arcot et al., (1995b) previously reported that

only about one fourth of the Alu elements containing $(AC)_n$ repeats had them as a part of their middle A-rich region. The one specific example they studied in detail had an evolutionary expansion of the A-rich region (orangutan and gibbon) before the genesis of the AC repeat; suggesting the requirement for an initial expansion. Interestingly, our large-scale analysis of the middle A-rich regions of Ya5 elements demonstrates a trend toward expansion of the A region, providing additional support for this region of the Alu elements to act as a potential nucleus for the genesis of simple sequence repeats.

From our subset of 269 AluYa5 elements, we were able to identify a new Alu subfamily termed Ya5a2. The estimated average age of 0.62 my (0.28 - 1.08 my with 95% confidence) makes Ya5a2 the youngest subfamily of Alu repeats identified in the human genome to date. It is as abundant as the Ya8 subfamily (Roy et al. 2000) and its higher level of insertion polymorphism suggests a higher level of current retroposition. The Ya5a2 subfamily may have originated from a Ya5 Alu element that inserted in a genomic region that favored transcription and corresponding retroposition activity of the element, generating a source gene. The subsequent accumulation of the two specific mutations facilitated the differentiation of the copies made by the Ya5a2 source gene from the larger background of several hundred genomic Ya5 Alu family members. As new Alu elements integrate into the genome in favorable genomic locations they can occasionally remain retropositionally competent and generate copies of themselves. However, the frequency of fortuitous insertions of new Alu elements into favorable genomic locations for subsequent mobilization is still a rare event since the continuity of the hierarchical subfamily sequence structure of the Alu elements is largely conserved throughout primate evolution.

Alu elements that are polymorphic for insertion presence/absence have previously proven useful for the study of human population genetics and forensics (Batzer et al. 1991; Batzer et al. 1994; Stoneking et al. 1997). The identification of a very young Alu subfamily with a high proportion of polymorphic members provides a new source of Alu insertion polymorphisms for the study of human population genetics. However, it is important to note that theYa5a2 subfamily is extremely small (about 35 copies total in a background of over 1,000,000) comparable to Ya8, so that an exhaustive analysis of a single human genome would only generate about 20 polymorphic Ya5a2 elements.

Since our analysis of Alu elements related to the Apert's insertion only included about 40% of the human genome (both finished and draft sequence included), there are possibly one or two other perfect complements in the human genome that have not yet been sequenced and may be the actual source gene for these elements. The transcriptional potential of this element would be consistent with its role as the potential source Alu gene. This confirms the existence of minor active source genes that differ from the source gene that generated almost all the Alu elements present in the human genome today. In addition, the *de novo* Apert's Alu element was also transcriptionally active. There are two possible explanations for this result. First, the transcriptional capacity of the elements was evaluated by transient transfections in tissue culture. This system does not reflect the influence of chromatin structure and methylation patterns (position effects) on the transcription and presumably retroposition potential of the two Alu repeats. Alternatively, the *de novo* Apert's Alu element may have inserted in a region of the FGFR2 gene that fortuitously enhances its own transcription capability. Although, further studies will be required to make more definitive statements in this

regard, the transcriptional capability of Ya5NBC237 is consistent with one of the many requirements a source gene possesses, making it a plausible candidate source gene for the *de novo* Apert's insertion.

In summary, the computational analyses of a subset of recently integrated Alu elements demonstrate that Alu sequence evolution is affected by a number of dynamic events. New retroposition competent Alu source genes, gene conversion, and genetic instability each play an important role in Alu sequence evolution and proliferation within the human genome.

MATERIALS AND METHODS

Computational analyses

Screening of the GenBank non-redundant (nr), the high throughput genome sequence (htgs) and the genomic survey sequence (gss) databases were performed using the Advanced Basic Local Alignment Search Tool 2.0 (BLAST) (Altschul et al. 1990) available from the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/). For the Ya5 subfamily analysis, the database was searched for matches to the 281 bases of the Ya5 consensus sequence with the following advanced options: -e 1.0 e-120, -b 1000, and –v 1000. A region composed of 500 bases of flanking DNA sequence directly adjacent to the sequences identified from the databases that matched the initial GenBank BLAST query were subjected to annotation using either RepeatMasker2 from the University of Washington Genome Center server (http://ftp.genome.washington.edu/cgi-bin/RepeatMasker) or Censor from the Genetic Information Research Institute (http://www.girinst.org/Censor_Server-Data_Entry_Forms.html) (Jurka et al. 1996). These programs annotate the repeat sequence content of DNA sequences from humans and rodents. The sequences were then subjected to more detailed analysis using MegAlign (DNAStar version 3.1.7 for Windows 3.2). The following parameters were used to select the Ya5 elements to be analyzed: 1- Ya5 had to have all 5 diagnostic nucleotides (except for the first position since it is a highly mutable CpG). 2- No truncated Alu elements were included in the analysis. 3- No Alu elements identified as a result of directed cloning strategies designed to identify Alu repeats were included (only those randomly found within larger data sequence). 4- Duplicate Alu elements were eliminated based on flanking sequences. The

consensus sequences of the Yb8 and Ya8 subfamilies were used for parallel searches of the three GenBank databases mentioned above. A complete list of the Alu elements identified from the GenBank search is available from MAB or PLD.

To search for putative source genes of the Alu elements that have previously been associated to different diseases, the three GenBank databases were searched using the sequence of each individual repeat to identify exact complements (Deininger and Batzer 1999 and references therein).

DNA samples

Human DNA samples from the European, African-American, Egyptian and Greenland Native population groups were isolated from peripheral blood lymphocytes (Ausubel et al. 1996) that were available from previous studies (Roy et al. 2000).

Oligonucleotide primer design and PCR amplification

A region composed of approximately 500 bases of flanking unique DNA sequences adjacent to each Alu repeat were used to design primers for fourteen Ya5a2 Alu elements (13 exact matches to consensus, Table 2). PCR primers were designed using the Primer3 software (Whitehead Institute for Biomedical Research, Cambridge, MA, USA) (http://www.genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi). The resultant PCR primers were screened against the GenBank non-redundant database for the presence of repetitive elements using the BLAST program, and primers that resided within known repetitive elements were discarded and new primers were designed. PCR amplification was carried out in 25 μl reactions using 50-100 ng of target DNA, 40 pM of each

oligonucleotide primer, 200 μM dNTPs in 50 mM KCl, 1.5 mM $MgCl_2$, 10 mM Tris-HCl pH 8.4 and Taq® DNA polymerase (1.25 U) as recommended by the supplier (Life Technologies). Each sample was subjected to the following amplification cycle: an initial denaturation of 2:30 min at 94°C, 1 min of denaturation at 94°C, 1 min at the annealing temperature, 1 min of extension at 72°C, repeated for 32 cycles, followed by a final extension at 72°C for 10 min. Twenty microliters of each sample was fractionated on a 2% agarose gel with 0.25μg/ml ethidium bromide. PCR products were directly visualized using UV fluorescence. The human genomic diversity associated with each element was determined by the amplification of 20 individuals from each of four populations (African American, Greenland Native, European and Egyptian; 160 total chromosomes). The chromosomal location for elements identified from randomly sequenced large-insert clones was determined by PCR analysis of NIGMS human/rodent somatic cell hybrid mapping panels 1 and 2 (Coriell Institute for Medical Research, Camden, NJ).

Construction of plasmids

The following constructs were made: p^{-416}Ya5NBC237 (416 bp upstream genomic – Alu – 223 bases downstream); p^{-290}Ya5Ap (290 bp upstream genomic – Alu – 293 bases); and p^{NP}Ya5NBC237 ("no upstream" vector flank –Alu – 223 bases). Unless otherwise noted, PCR was performed in 20 μl reactions using an MJ Research PTC 200 thermal cycler with the following conditions: 1X Promega buffer, 1.5 mM $MgCl_2$, 200 μM dNTPs, 0.25 μM primers, 1.5 U *Taq* polymerase (Promega) at 94°C-2 min; 94°C-20 sec, 55°C (annealing temperature)-20 sec, 72°C-1 min, for 30 cycles; 72° C - 3 min. To PCR

amplify and clone the 864 bp fragment containing the *de novo* Alu Ya5 from Apert syndrome patient 1 (Accession # AF097344) the following primers were used: Forward 5'-GGTGTGGCCAAAGTGGAGGATGTGTAC-3' and Reverse 5'-TTATTCAAGGATAAAAGGGGCCATTTC-3' with an annealing temperature of 50°C; and for the 920 bp fragment containing AluYa5NBC237 (Accession # AL031274) the primers used were: Forward 5'- TTATTCCATTGGTCCTTTCCACCAG-3' and R 5'-CAGGCAGGGAGGTACTTGTCTCTTG-3' with an annealing temperature of 55°C. For the p^{NP}Ya5NBC237 PCR amplification from the clone was done using the same reverse primer and the FAlu5 primer: 5'-GGCCGGGCGCGGTGGCTCA-3'.

The final PCR product of the complete construct was cloned into pGEMTeasy Vector System I (Promega). Constructs were subjected to DNA sequence analysis in order to verify their sequence context. Purified plasmids from the constructs were prepared by alkaline lysis of bacterial cells followed by banding in a CsCl gradient twice. DNA concentrations were determined spectrophotometrically by using A_{260} and verified by visual examination of ethidium bromide-stained agarose gels.

Alu Transcription in Cell Lines and RNA Analysis

Transient transfections were carried out in the rodent cell line C6 glioma (ATCC CCL107). Monolayers were grown to 50-70% confluency and transfected with 3μg of the construct-containing plasmid and 1μg of control plasmid (p^{7SL}BC1) using LipofectAmine Plus® (Gibco Life Sciences) following the manufacturer's recommended protocol. Total RNA was isolated 16 – 20 h post-transfection.

RNA was extracted from cell lines utilizing the Trizol™ Reagent (Life Technologies, Inc.) according to the manufacturer's protocol. Equal amounts of RNA were fractionated on a 2% agarose-formaldehyde gel and then transferred to a nylon membrane, Hybond-N (Amersham). Northern blots were hybridized utilizing the following end-labeled oligonucleotide probes: unique-1 5'-TGTGTGTGCCAGTTACCTTG-3' (complementary to 3' end of the control plasmid) and AluYA5-1 5'-ACCGTTTTAGCCGGGAATGGTC-3' (complementary to Ya5 Alu RNA, but not to 7SL) in 5X SSC, 5X Denhardt's, 1% SDS and 100 μg/ml herring sperm DNA. Oligonucleotides were end-labeled by incorporating [γ-^{32}P] ATP (Amersham) with T4 polynucleotide kinase (New England BioLabs), and subsequently separated from free label by filtration through a Sephadex G-50 column. Blots were washed three times at 45°C with a low stringency buffer (2X SSC and 1% SDS) and subjected to autoradiography or quantified using a FujiFilm FLA-2000 fluorescent image analyzer (Fuji Photo Film Co., LTD, Tokyo, Japan). Statistical analysis was performed using the Jandel SigmaStat Statistical Software Version 2, Jandel Corporation.

ACKNOWLEDGMENTS

AMR was supported by a Brown Foundation fellowship from the Tulane Cancer Center. This research was supported by National Institutes of Health RO1 GM45668 to PLD, Department of the Army DAMD17-98-1-8119 to PLD and MAB, and award number 1999-IJ-CX-K009 from the Office of Justice Programs, National Institute of Justice, Department of Justice to MAB. Points of view in this document are those of the authors and do not necessarily represent the official position of the U.S. Department of Justice.

REFERENCES

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J.Mol.Biol.* **215:** 403-410.

Arcot, S.S., T.H. Shaikh, J. Kim, L. Bennett, M. Alegria-Hartman, D.O. Nelson, P.L. Deininger, and M.A. Batzer. 1995a. Sequence diversity and chromosomal distribution of "young" Alu repeats. *Gene* **163:** 273-278.

Arcot, S.S., Z. Wang, J. Weber, P.L. Deininger, and M.A. Batzer. 1995b. *Alu* repeats: A source for the genesis of primate microsatellites. *Genomics* **29:** 136-144.

Ausubel, F.M., R. Brent, R.E. Kingston, D.D. Moore, J.G. Seidman, J.A. Smith, and K. Struhl. 1996. *Current Protocols In Molecular Biology*, John Wiley & Sons, Inc. Canada.

Batzer, M.A., P.L. Deininger, U. Hellmann-Blumberg, J. Jurka, D. Labuda, C.M. Rubin, C.W. Schmid, E. Zietkiewicz, and E. Zuckerkandl. 1996a. Standardized nomenclature for *Alu* repeats. *J.Mol.Evol.* **42:** 3-6.

Batzer, M.A., S.S. Arcot, J.W. Phinney, M. Alegria-Hartman, D.H. Kass, S.M. Milligan, C. Kimpton, P. Gill, M. Hochmeister, P.A. Ioannou, R.J. Herrera, D.A. Boudreau, W.D. Scheer, B.J. Keats, P.L. Deininger, and M. Stoneking. 1996b. Genetic variation of recent Alu insertion in human populations. *J.Mol.Evol.* **42:** 22-29.

Batzer, M.A., C.M. Rubin, U. Hellmann-Blumberg, M. Alegria-Hartman, E.P. Leeflang, J.D. Stern, H.A. Bazan, T.H. Shaikh, P.L. Deininger, and C.W. Schmid. 1995.

Dispersion and insertion polymorphism in two small subfamilies of recently amplified human Alu repeats. *J.Mol.Biol.* **247:** 418-427.

Batzer, M.A., M. Stoneking, M. Alegria-Hartman, H. Bazan, D.H. Kass, T.H. Shaikh, G.E. Novick, P.A. Ioannou, W.D. Scheer, R.J. Herrera, and P.L. Deininger. 1994. African origin of human-specific polymorphic Alu insertions. *Proc.Natl.Acad.Sci.,USA* **91:**12288-12292.

Batzer, M.A., V. Gudi, J.C. Mena, D.W. Foltz, R.J. Herrera, and P.L. Deininger. 1991. Amplification dynamics of human-specific (HS) Alu family members. *Nucleic Acids Res.* **19:** 3619-3623.

Batzer, M.A., G. Kilroy, P.E. Richard, T.H. Shaikh, T. Desselle, C. Hoppens, and P.L. Deininger. 1990. Structure and variability of recently inserted Alu family members. *Nucleic Acids Res.* **18:** 6793-6798.

Belmaaza, A., J.C. Wallenburg, S. Brouillette, N. Gusew, and P. Chartrand. 1990. Genetic exchange between endogenous and exogenous LINE-1 repetitive elements in mouse cells. *Nucleic.Acids.Res.* **18:** 6385-6391.

Bird, A.P. 1980. DNA methylation and the frequency of CpG in animal DNA. *Nucleic.Acids.Res.* **8:** 1499-1504.

Deininger, P.L., M.A. Batzer, I.C. Hutchison, and M. Edgell. 1992. Master genes in mammalian repetitive DNA amplification. *Trends in Genetics* **8:** 307-312.

Deininger, P.L., and M.A. Batzer. 1993. Evolution of Retroposons. In *Evolutionary Biology* (eds. Heckht, M.K. and et al) , pp. 157-196. Plenum Publishing, New York.

Deininger, P.L., and G. Daniels. 1986. The recent evolution of mammalian repetitive DNA elements. *Trends in Genetics* **2:** 76-80.

Deininger, P.L. and M.A. Batzer. 1999. Alu repeats and human disease. *Mol.Genet.Metab.* **67:** 183-193.

Dombroski, B.A., A.F. Scott, and H.H. Kazazian Jr. 1993. Two additional potential retrotransposons isolated from a human L1 subfamily that contains an active retrotransposable element. *Proc.Natl.Acad.Sci.U.S.A.* **90:** 6513-6517.

Economou, E.P., A.W. Bergen, A.C. Warren, and S.E. Antonarakis. 1990. The polydeoxyadenylate tract of Alu repetitive elements is polymorphic in the human genome. *Proc.Natl.Acad.Sci., USA* **87:** 2951-2954.

Hoff, E.F., H.L. Levin, and J.D. Boeke. 1998. Schizosaccharomyces pombe retrotransposon Tf2 mobilizes primarily through homologous cDNA recombination. *Mol.Cell Biol.* **18:** 6839-6852.

Jelinek, W.R., and C.W. Schmid. 1982. Repetitive sequences in eukaryotic DNA and their expression. *Annu.Rev.Biochem.* **51:** 813-844.

Jurka, J., D.J. Kaplan, C.H. Duncan, J. Walichiewicz, A. Milosavljevic, G. Murali, and J.F. Solus. 1993. Identification and characterization of new human medium reiteration frequency repeats. *Nucleic.Acids.Res.* **21:** 1273-1279.

Jurka, J., P. Klonowski, V. Dagman, and P. Pelton. 1996. CENSOR—a program for identification and elimination of repetitive elements from DNA sequences. *Comput. Chem.* **20:** 119-121.

Jurka, J., and T. Smith. 1988. A fundamental division in the Alu family of repeated sequences. *Proc. Natl. Acad. Sci., U.S.A.* **85:** 4775-4778.

Jurka, J. and C. Pethiyagoda. 1995. Simple repetitive DNA sequences from primates: compilation and analysis. *J Mol.Evol.* **40:** 120-126.

Jorde, L.B., W.S. Watkins, M.J. Bamshad, M.E. Dixon, C.E. Ricker, M.T. Seielstad, and M.A. Batzer. 2000. The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. *Am.J.Hum.Genet.* **66**: 979-988.

Kapitonov, V., and J. Jurka. 1996. The age of Alu subfamilies. *J.Mol.Evol.* **42:** 59-65.

Kass, D.H., M.A. Batzer, and P.L. Deininger. 1995. Gene conversion as a secondary mechanism in SINE evolution. *Mol.Cell.Biol.* **15:** 19-25.

Labuda, D., and G. Striker. 1989. Sequence conservation in Alu evolution. *Nucleic.Acids.Res.* **17:** 2477-2491.

Lester, T., C. McMahon, N. VanRegemorter, A. Jones, and S. Genet. 1997. X-linked immunodeficiency caused by insertion of Alu repeat sequences. *J Med.Gen.Suppl.* **34:** S81.

Miki, Y., T. Katagiri, F. Kasumi, T. Yoshimoto, and Y. Nakamura. 1996. Mutation analysis in the BRCA2 gene in primary breast cancers. *Nat.Genet.* **13:** 245-247.

Miyamoto, M.M., J.L. Slightom, and M. Goodman. 1987. Phylogenetic relations of humans and African apes from DNA sequences in the psi eta-globin region. *Science* **238:** 369-373.

Montermini, L., E. Andermann, M. Labuda, A. Richter, M. Pandolfo, F. Cavalcanti, L. Pianese, L. Iodice, G. Farina, A. Monticelli, M. Turano, A. Filla, M.G. De, and S. Cocozza. 1997. The Friedreich ataxia GAA triplet repeat: premutation and normal alleles. *Hum.Mol.Genet.* **6:** 1261-1266.

Oldridge, M., E.H. Zackai, D.M. McDonald-McGinn, S. Iseki, G.M. Morriss-Kay, S.R. Twigg, D. Johnson, S.A. Wall, W. Jiang, C. Theda, E.W. Jabs, and A.O. Wilkie. 1999. De novo Alu-element insertions in FGFR2 identify a distinct pathological basis for Apert syndrome. *Am.J.Hum.Genet.* **64:** 446-461.

Perna, N.T., M.A. Batzer, P.L. Deininger, and M. Stoneking. 1992. Alu insertion polymorphism: a new type of marker for human population studies. *Hum.Biol.* **64**: 641-648.

Rogers, J.R., and K.R. Willison. 1983. A major rearrangement in the H-2 complex of mouse t haplotypes. *Nature* **304:** 549-552.

Roy, A.M., M.L. Carroll, D.H. Kass, S.V. Nguyen, A-H. Salem, M.A. Batzer, and P.L. Deininger. 2000. Recently integrated human Alu repeats: finding needles in the haystack. *Genetica* **107**:1-13.

Schmid, C.W., and R. Maraia. 1992. Transcriptional regulation and transpositional selection of active SINE sequences. *Curr.Opin.Genet.Dev.* **2:** 874-882.

Shaikh, T.H., and P.L. Deininger. 1996. The role and amplification of the HS Alu subfamily founder gene. *J.Mol.Evol.* **42:** 15-21.

Shaikh, T.H., A.M. Roy, J. Kim, M.A. Batzer, and P.L. Deininger. 1997. cDNAs derived from primary and small cytoplasmic Alu (scAlu) transcripts. *J Mol.Biol.* **271:** 222-234.

Shen, M., M.A. Batzer, and P.L. Deininger. 1991. Evolution of the Master Alu Gene(s). *J.Mol.Evol.* **33:** 311-320.

Sinnett, D., C. Richer, J.M. Deragon,. and D. Labuda. 1992. Alu RNA transcripts in human embryonal carcinoma cells. Model of post-transcriptional selection of master sequences. *J.Mol.Biol.* **226**, 689-706.

Slagel, V., E. Flemington, V. Traina-Dorge, H. Bradshaw, and P.L. Deininger. 1987. Clustering and sub-family relationships of the Alu family in the human genome. *Mol.Biol.Evol.* **4:** 19-29.

Smit, A.F. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* **9:** 657-663.

Stoneking, M., J.J. Fontius. S.L. Clifford, H. Soodyall, S.S Arcot, N. Saha, T. Jenkins, M.A. Tahir, P.L. Deininger, and M.A. Batzer. 1997. Alu insertion polymorphisms and human evolution: evidence for a larger population size in Africa. *Genome Res.* **7:** 1061-1071.

Tishkoff, S.A., G. Ruano, J.R. Kidd, and K.K. Kidd. 1996. Distribution and frequency of a polymorphic Alu insertion at the plasminogen activator locus in humans. *Hum.Genet.* **97**: 759-764.

Wallace, M.R., L.B. Andersen, A.M. Saulino, P.E. Gregory, T.W. Glover, and F.S. Collins. 1991. A de novo Alu insertion results in neurofibromatosis type 1. *Nature* **353:** 864-866.

Weiner, A., P.L. Deininger, and A. Efstradiatis. 1986. The Reverse Flow of Genetic Information: pseudogenes and transposable elements derived from nonviral cellular RNA. *Annual Reviews of Biochemistry* **55:** 631-661.

Willard, C., H.T. Nguyen, and C.W. Schmid. 1987. Existence of at least three distinct Alu subfamilies. *J.Mol.Evol.* **26:** 180-186.

FIGURE LEGENDS

Fig. 1. Consensus sequence alignment of Ya5, and the potential new subfamily members identified. Nucleotide substitutions at each position are indicated with the appropriate nucleotide. Deletions are marked by dashes (-). The Ya5 diagnostic nucleotides are indicated in bold with the corresponding diagnostic number above as defined by (Shen et al. 1991).

Fig. 2. Schematic for the evolution of recently integrated Alu subfamilies. The origin of the Ya5a2 Alu subfamily is shown after the divergence of Ya5 and Yb8 elements. The total number of elements found in the nr-database (perfect matches in parenthesis) are shown first separated by a slash from the total number of elements found in all three databases (nr, gss, htgs). For the Ya5 elements only the nr-database results are shown.

Fig. 3. Evolution of the diagnostic nucleotide positions from Y to Ya5 Alu elements. Alignment of the five Alu Ya5 diagnostic nucleotides as defined by (Shen et al. 1991) and the different Ya1, 2, 3, 4 elements found in the nr-database. For easy reference, individual elements containing different combinations of the diagnostic mutations were numbered consecutively in order of abundance (Ya3.1, Ya3.2 , etc.). Ya4.4 elements were considered as Ya5 elements in the first Ya5 subfamily analysis in this paper. The total number of elements found for each subgroup is indicated on the left in parenthesis. Potential forward (f) or backward (b) gene conversions are indicated on the right. The previously reported order of appearance of Ya5 diagnostic mutations (Shaikh and

Deininger 1996) is indicated below. Elements with diagnostic mutations that follow the stepwise hierarchical accumulation are circled.

Fig. 4. Evaluation of transcriptional capability of the potential FGFR2 source Ya5 Alu element. The transcriptional efficiency of the *de novo* FGFR2 Alu repeat and its putative source gene were evaluated by northern blot analysis from transient transfection studies. The following constructs were evaluated: 1- p^{-290}Ap, 2- p^{-416}Ya5NBC237 and p^{NP}Ya5NBC237. Lane 4 and 5 are internal control only and no DNA controls respectively. Small arrows indicate the Alu transcripts and the open arrow indicates the internal control transcript. The ratio of the Alu transcript/ control transcript (numbers below) was normalized to the p^{NP}Ya5NBC237 transcription ratio, which was assigned the arbitrary value of 1.

Figure 1

Alu middle A-rich region

Ya5 middle A rich region	**A_n**							
	4	**5**	**6**	**7**	**8**	**9**	**10**	**11**
T(A_n)TACA$_6$TT [1]	0	269 [3]	9	1	0	1	-	-
TA$_5$TAC(A_n)TT [2]	0	2	269 [3]	37 [4]	11	7	3	0

[1] n = 5 in Ya5 consensus
[2] n = 6 in Ya5 consensus
[3] Data from the non-redundant database only
[4] All 23 Ya5a2 members are included

Figure 2

Alu Ya5a2 PCR Primers, Chromosomal Locations, and PCR Product Sizes

Name	5' Primer sequence (5'-3')	3' Primer sequence (5'-3')	A.T.[1]	Chromosomal[2] Location	Product Size[3] Filled	Empty
Ya5NBC206	TCCTTAGCTATCTCACAAGCTACAT	ACACATTTCCTTCAAGAGGTCAAAG	60°C	4	734	424
Ya5NBC207	CAGTTTTATACACTGGCCTGTTTTC	TTGTAGGAGAAAGAGGGGAAATACT	50°C	6	443	122
Ya5NBC208	AATACCTTGTACATCTTCACCCCTA	TCTCTCTGCTGCACAGTTTGTT	50°C	14	441	115
Ya5NBC240	CAGGAGATAAATATGTTCGGAGAGT	TAACTGGGACAGTGAGTTTTACCTG	55°C	9	505	202
Ya5NBC241	GGTTCCAATAGAGAGCAACAGAA	ACCTTAAGCTTTCCCCCAGA	55°C	15	392	66
Ya5NBC242	AACAAAATTCCCTTTCCTCCA	GGCAATCTGACCTTGGGTAA	55°C	7	503	192
Ya5NBC7	TGATGGATATTTGGGTTGGTTC	GGACTGTAAACTAGTTCAACCATTGTG	60°C	7	522	216
Ya5NBC205	ACATGAAGGGCCGACTGTAT	TGCTGCTGCATTATCAACTG	50°C	21	435	81
Ya5NBC209	GTCTATGGGAAGATGAAGAATAGGA	GATGGAGTCACTCATGTGAAAAGTA	55°C	14	447	116
Ya5NBC239	CAGCTGAGAACTGTCACAAATAGAA	ATCAATGACTGACTTGTGCTGAGT	55°C	9	531	198
Ya5NBC243	CCATGATTCGTCATTCACCA	AGGAGACCTGCCAATGAATG	60°C	21	406	86
Ya5NBC220	AAATCAAGCTGCCATACCTCA	GAAACCATCCTTCACAGTGG	60°C	1	463	141
Ya5NBC235	CCCAAGGCACTTGCTGTTA	CCCTTCGAGAAAGAGGAAGG	50°C	2	391	76
Ya5NBC244	CCTATGGCTGAAATCTTCTGAAACT	ATATCTTGGTCCACTAGACAAGCAC	60°C	18	453	130
Ya5NBC237*	CCCATGGAGGGTCTTTCCTA	CTGGAAACCATCCTTCACAGT	60°C	1	410	88

1. Amplification of each locus required 2:30 min @ 94°C initial denaturing, and 32 cycles for 1min 94°C, 1 min Annealing Temperature (A.T.) and 1 min elongation at 72°C. A final extension time of 10 min at 72°C was also used.
2. Chromosomal location determined from Accession information or by PCR analysis of NIGMS monochromosomal hybrid cell line DNA samples.
3. Empty product sizes calculated by removing the Alu element and 1 direct repeat from the filled sites that were identified.

* Alu Ya5a2 element of the $FGFR_2$ gene.

Figure 3

Alu Ya5a2 (*NF1*) associated human genomic diversity.

Ya5a2 elements	Accession # (duplicates)	Position	Allele frequency[1]
Ya5NBC206	AC004057	76767-77048	fixed present
Ya5NBC207	AL118555 (AL132992)	9981-9700 (40728-41009)	fixed present
Ya5NBC208	AL109919	70170-69889	intermediate
Ya5NBC220	AC007611	136715-136434	intermediate
Ya5NBC240	AC133410 (AL135841)	34800-35081 (49829-49548)	intermediate
Ya5NBC241	AC018924	144017-144298	intermediate
Ya5NBC242	AC009517	161301-161582	intermediate
Ya5NBC7	AC004848	24522-24241	low
Ya5NBC205	AL011328	204488-204207	low
Ya5NBC209	AC00808	147056-146775	low
Ya5NBC239	AL133284	115867-115586	low
Ya5NBC244	AC026839	64885 - 64604	low
Ya5NBC243	AJ011929	151192-151473	low
Ya5NBC235[2]	AQ748733	458-739	fixed present
Ya5NBC237[3]	AL031274	33175-33501	intermediate

[1] Allele frequency was classified as: fixed present, fixed absent, low, intermediate, or high frequency insertion polymorphism. Fixed present: every individual tested had the Alu element in both chromosomes. Low frequency insertion polymorphism: the absence of the element from all individuals tested, except for one or two homozygous or heterozygous individuals. Intermediate frequency insertion polymorphism: the Alu element is variable as to its presence or absence in at least one population. High frequency insertion polymorphism: the element is present in all individuals in the populations tested, except for one or two heterozygotous or absent individuals.

[2] several N's

[3] Ya5NBC237 is the exact match to the FGFR2 Alu insertion.

Figure 4

Alu Ya5a2 Associated Human Genomic Diversity

Elements	African American				Greenland Natives				European				Egyptian				
	Genotypes[1]			*f*Alu[2]	Genotypes			*f*Alu	Genotypes			*f*Alu	Genotypes			*f*Alu	Het.[3]
Ya5NBC20	20	0	0	1.000	20	0	0	1.000	20	0	0	1.000	20	0	0	1.000	0.000
Ya5NBC20	20	0	0	1.000	20	0	0	1.000	20	0	0	1.000	20	0	0	1.000	0.000
Ya5NBC20	4	1	7	0.375	3	0	4	0.429	13	0	6	0.684	7	0	5	0.583	0.482
Ya5NBC23	5	6	2	0.615	5	8	6	0.474	15	5	0	0.875	6	8	1	0.667	0.422
Ya5NBC24	5	1	9	0.367	11	0	4	0.733	5	1	10	0.344	5	3	3	0.591	0.464
Ya5NBC24	3	9	5	0.441	6	11	2	0.605	0	7	11	0.194	3	8	4	0.467	0.459
Ya5NBC24	2	13	1	0.531	7	4	3	0.643	3	4	11	0.278	3	3	1	0.643	0.474
Ya5NBC7	0	0	19	0.000	0	0	20	0.000	0	0	20	0.000	0	0	20	0.000	0.000
Ya5NBC20	0	0	20	0.000	0	0	20	0.000	0	0	20	0.000	0	0	20	0.000	0.000
Ya5NBC209	0	1	17	0.028	0	0	17	0.000	0	0	19	0.000	0	0	19	0.000	0.000
Ya5NBC239	0	0	20	0.000	0	0	20	0.000	0	0	20	0.000	0	0	20	0.000	0.000
Ya5NBC243	0	0	20	0.000	0	0	20	0.000	0	0	20	0.000	0	0	20	0.000	0.000
Ya5NBC220	0	14	5	0.368	1	15	2	0.472	0	18	1	0.474	0	9	2	0.409	0.502
Ya5NBC244	0	0	12	1.000	-	-	-	-	0	0	10	0.000	0	0	8	0.000	0.000
Ya5NBC235	20	0	0	1.000	20	0	0	1.000	20	0	0	1.000	20	0	0	1.000	0.000
Ya5NBC237*	18	1	0	0.974	15	4	0	0.895	20	0	0	1.000	18	1	0	0.974	0.075

1. Genotypes: +/+ Alu, +/- Alu, -/- Alu
2. Frequency of the presence of the Alu
3. Average heterozygosity
- not determined
* Ya5NBC237 is the exact match to the FGFR2 Alu insertion